The Dissertation Committee for Ben Shogo Wendel certifies that this is the approved version of the following dissertation:

ANALYZING INFECTION-DRIVEN IMMUNE PERTURBATIONS BY QUANTITATIVE IR-SEQ

Committee:

Ning Jiang, Co-Supervisor

George Georgiou, Co-Supervisor

Hal Alper

Mark Davis

Jennifer Maynard

ANALYZING INFECTION-DRIVEN IMMUNE PERTURBATIONS BY QUANTITATIVE IR-SEQ

by

Ben Shogo Wendel

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2017

Dedication

To my grandmother, Mitsue, who inspired me to study human health and taught me to never give up, no matter the odds. And to the rest of my family, thank you for your unconditional love and support.

Acknowledgements

I would like to thank my fellow Jiang Lab members for all of your help throughout the years. Shuqi Zhang, for accompanying me on this journey from the very beginning and never failing to provide help and advice when I needed it most. Chenfeng He, for showing me the ropes of bioinformatics analysis and always taking the time to help analyze data without hesitation. Keyue Ma, for allowing me to pick your brain about molecular biology and the countless times you helped me with PCR-related experiments. Stefany Hernandez, for all of the time and effort you put into helping me prepare sequencing libraries. Brittain Sobey, Michael Don, and Kate Baird, for helping me keep my head above water from an administrative standpoint – easier said than done. Di Wu, Chad Williams, Evan Cohan, Mingjuan Qu, Qian Shi, Yang Liu, Alex Schonnesen, Tian Li, Mary Salazar, and everyone else in the lab, for all of your help along the way.

I would like to acknowledge George Georgiou, Erik Johnson, and Wissam Charab, for everything from teaching me to culture the ever-finicky HEK293 cells to brainstorming applications of immune repertoire sequencing. Susan Pierce, Peter Crompton, and Eugene Liu, for providing samples and guidance for malaria-related experiments. Laura Su and Daniel del Alcazar, for gathering samples, analyzing CyTOF data, and providing indispensable guidance for the HIV project. Jessica Wheeler Podnar and the rest of the Genome Sequencing and Analysis Facility and Jeremy Day, for organizing and running endless sequencing runs. Mark Davis, Wong Yu, and Natalia Sigal, for teaching me how to perform tetramer enrichments, helping me produce a vast quantity of MHC monomers, and all the guidance along the way. Hal Alper and Jennifer Maynard, for all of your guidance and direction. Pengyu Ren, for providing the

computational resources to make all of the sequencing analysis possible. Keke Chen and Jun Xiao, for all of your help with antibody lineage tree construction.

Finally, I would like to thank my advisor, Jenny Jiang, for all of the advice, direction, guidance, and support; for kicking me in the butt when I needed it and talking me off the ledge when things went wrong. Without you, none of this would have been possible.

Thank you to everyone who helped make my graduate studies an unforgettable experience.

ANALYZING INFECTION-DRIVEN IMMUNE PERTURBATIONS BY QUANTITATIVE IR-SEQ

Ben Shogo Wendel, Ph.D.

The University of Texas at Austin, 2017

Supervisor: Ning Jiang

Immune repertoire sequencing (IR-Seq) rapidly emerged with the advent of high-throughput sequencing as a means of characterizing the adaptive immune system. Early generations of IR-Seq were plagued by sequencing errors and low diversity coverage. We developed Molecular Identifier Clustering-based IR-Seq (MIDCIRS) to quantitatively and comprehensively measure the immune repertoire from a small amount of blood. We used naive B cells to formulate a general framework for IR-Seq experimental validation and quality control and showed that MIDCIRS can be applied to as few as 1,000 naive B cells with excellent diversity coverage.

Using MIDCIRS, we studied the antibody repertoire response to acute malaria infection in young children. We found that the infant antibody repertoire is surprisingly competent at introducing somatic hypermutations (SHM) and diversifying B cell clonal lineages in response to malaria infection. Detailed analysis of memory B cell-containing lineages in malaria-experienced toddlers revealed that memory B cells further mutate upon malaria rechallenge. IgM-expressing memory B cells largely retain IgM expression upon rechallenge, but a subset class switch to IgG and IgA.

Accurate antibody repertoire analysis requires not only accurate sequencing data, but also correct reference germline allele sequences. Mismatches between the reference sequence and an individual's true germline sequence would be mistakenly counted as SHMs, inflating the SHM load and skewing the repertoire analysis. We developed a simple yet effect method for predicting novel germline allele sequences from antibody repertoire data and validating them via targeted sequencing of non-rearranged genomic DNA.

HIV infection has a profound impact on the $CD4^+$ T cell compartment, which can have a devastating effect on the adaptive immune system as a whole. Paradoxically, while peripheral $CD4^+$ T cell counts drop with disease severity, T_{FH} cells show an inverse relationship. We found that these expanded T_{FH} cells exhibit a functionally restricted phenotype, which could contribute to ineffective antibody responses both to HIV and unrelated vaccines. Using MIDCIRS, we found that these expanded T_{FH} cells are enriched with HIV-specific sequences and show signs of antigen-driven convergent evolution, suggesting that HIV-specific T cells are selected and recruited into the T_{FH} compartment during infection.

Table of Contents

List of Tables

List of Figures

Figure 3.8: Antigen selection strength comparisons between infants and toddlers. Selection strength distributions, as determined by BASELINe[25], were compared between infants (black) and toddlers (red) for PBMCs from pre-malaria (**a-c**) (N_{infant}=6, $N_{toddler}$=9) and acute malaria (**d-f**) (N_{infant}=6, $N_{toddler}$=9) timepoints, separated by isotype: (**a,d**) IgM, (**b,e**) IgG, and (**c,f**) IgA. Selection strength on CDR (CDR1 and 2, top half of each panel) and FWR (FWR2 and 3, bottom half of each panel) for unique RNA molecules was calculated. CDR3 and FWR4 were omitted due to the difficulty in determining the germline sequence. FWR1 for all sequences was also omitted because it was not covered entirely by some of the primers. *P* value calculated as previously described[25]. 46

Figure 5.1: Summary of experimental design. Cryopreserved LN samples were obtained from 7 healthy controls, 4 IBD patients, and 25 HIV^+ individuals. (**A**) Cells from all donors were stimulated with PMA plus ionomycin and analyzed on CyTOF. (**B**) LN cells from 8 HIV^+ donors were sorted by naïve ($CD45RO^-CXCR5^-CCR7^+CD27^+$), memory ($CD45RO^+CXCR5^-PD1^-ICOS^-$), or GC T_{FH} phenotype ($CD45RO^+CXCR5^+PD1^+CD57^+$) for TCR sequencing. (**C**) A subset of these donors also had enough cells for peptide stimulation in culture (5 for Gag and 6 for HA). After 3-4 weeks, cultured cells were restimulated with peptides and sorted for activation by CD69 and CD40L expression. TCR sequences obtained from Gag or HA-peptide reactive T cells were used as a reference sequence dataset to identify matching HA- or Gag-specific T cells from sorted and sequenced bulk populations from (**B**). Figure produced by L.F.S.....93

Figure 5.3: Cellular heterogeneity of GC T_{FH} cells across HC, IBD, and HIV patient-derived LNs. (**A**) Representative gates used to identify GC T_{FH} cells for t-SNE analysis. (**B**) All GC T_{FH} cells from 36 samples were concatenated and displayed on a two-dimensional t-SNE map. Manual gating was performed to identify t-SNE clusters based on contour map (left). Each cluster was assigned an arbitrary number and overlaid onto the t-SNE map (right). (**C**) Heatmap shows raw staining intensity of each marker within each cluster defined in (**B**) after arcsinh transformation. Staining intensities for each marker are shown for memory and naïve cells at the bottom of the heatmap for comparison. (**D**) Heatmap showing the frequency of each cluster in GC T_{FH} cells pooled by disease categories. Cluster groups, g1-g4, are defined by the dendogram (right), which was generated by hierarchical clustering based on cluster frequency in each sample type. (**E**) Stacked bar chart showing normalized frequency of phenotypic group distribution for each sample. HIV^+ samples were ordered by increasing $CD4^+$ T cell count (7 – 1136). Data and figure produced by L.F.S.

Figure 5.4: IL-21-secreting T_{FH} cells in HIV^+ patients acquire an activated phenotype and restricted functional diversity. (**A,B**) The number and frequency of $IL21^+$ T_{FH} cells as a percentage of cytokine producing GC T_{FH} cells. The denominator includes GC T_{FH} cells that produce any combination of IL-2, IFN-γ, TNF-α, IL-4, IL-21, and granzyme A. (**C,D**) Pie charts and bar graphs summarizing the frequency of $IL\text{-}21^+$ T_{FH} cells that produce only IL-21 or IL-21 plus 1, 2, 3, or 4 other effector molecules (IL-2, IFN- γ, TNF- α, IL-4, or granzyme A) in the GC subset (**C**) or $CXCR5^+CD45RO^+CD4^+$ T cells (**D**) from HC or HIV^+ LNs. (**E,F**) Correlation between switched memory B cell frequency and the frequency of IL-21 only or IL21+2 producing $CXCR5^+CD45RO^+$ cells. Data include all LNs (See also **Figure D.4**). (**G,H**) The frequency of $CD38^+$ or $Ki67^+$ GC T_{FH} cells. (**I**) Correlation between the frequency of $CD57^+PD1^+$ cells and $Ki67^+$ GC T_{FH} frequency. Data include all LNs. Statistical analysis using Student's t-test was corrected for multiple comparisons using Holm-Sidak method, with alpha=5%. Error bars represent SEM. Association is measured by Spearman rank correlation and least squares fit regression. * $P < 0.05$, ** $P < 0.005$; *** $P < 0.0005$. Data and figure produced by L.F.S. ..102

Figure 5.6: Antigen-driven clonal selection signature in GC T_{FH} cells of HIV^+ LNs. (**A**) Representative degeneracy plot from sample H2. Coding degeneracy level (number of unique TCR nucleotide (nt) sequences encoding a common CDR3 amino acid (aa) sequence) of each CDR3 aa sequence is plotted against their frequency (measured as % of total TCR transcript) in naïve, memory, and GC T_{FH} cells. Each dot is a unique CDR3 aa sequence. Red dashed lines indicate cutoffs for degenerate (2 or more nt sequences coding for the same aa sequence, horizontal) and expanded (0.1% or more of TCR transcripts, vertical) clones. Each panel is broken into 4 quadrants: Q1: degenerate-abundant clones; Q2: degenerate-rare clones; Q3: nondegenerate-rare clones; Q4: nondegenerate-abundant clones. Red arrow points to example degenerate clone in (**B**). (**B**) An example of CDR3 aa degeneracy. aa (top row) and nt (bottom row) sequences for each of 3 distinct nt sequences (0.41% of total TCR transcripts) that code for the same aa sequence as the arrow points in (**A**) with Y=3, X=0.41%. Red boxes and highlights indicate redundant codons. (**C**) Comparison of Q1 degenerate-abundant clone percentage in naïve (black), memory (blue), and GC T_{FH} (orange) cells. Grey lines connect samples from the same patient. Bars indicate means. $*P < 0.05$ by two-tailed Wilcoxon signed-rank test.

Figure 5.7: GC T_{FH} cells exhibit HIV-antigen-driven clonal expansion and selection. (**A**) Gag-specific TCR clones overlap with HIV^+ LN $CD4^+$ T cell populations. Each thin slice of the arc represents a unique TCR sequence, ordered by the clone size (darker green for larger clones, inner circle). Grey curves indicate Gag-specific TCR clones (nt sequences) found in naïve (black, outer circle), memory (blue, outer circle), and GC T_{FH} (orange, outer circle) populations. (**B**) Number of Gag-specific TCR clones observed in naïve (black), memory (blue), and GC T_{FH} (orange) populations. Grey lines connect samples from the same patient. Bars indicate means. (**C**) Mean clone size of Gag-specific clones (magenta) and clones of unknown specificity (black) from naïve, memory, and GC T_{FH} populations. (**D**) Degeneracy, or the number of distinct nt sequences per CDR3 aa sequence, of Gag-specific clones (magenta) and clones of unknown specificity (black) from naïve, memory, and GC T_{FH} populations. Data from all 5 subjects were aggregated for **C** and **D**. Error bars indicate SEM. ** $P < 0.01$, *** $P < 0.001$ by two-tailed Student's t-test.....110

Chapter 1: Background

1.1 Adaptive immunity

The immune system is necessarily complex to combat the universe of possible pathogens that an individual can be exposed to throughout his or her lifetime. The bulk of this complexity resides in the adaptive branch – the component that recognizes and responds to specific antigens. As its name implies, the adaptive immune system evolves to more effectively subdue foreign agents that it has previously encountered. Achieving this is no small task, and the underlying biology is a subject of great scientific interest, particularly to the medical field. Vaccines have long been used with great success to elicit an adaptive immune response in place of an actual infection to confer protection against diseases. More recently, therapeutic antibodies and engineered T cells have emerged as exciting areas of research towards the treatment of a wide range of diseases, including HIV[1], pertussis[2], Ebola[3], and various types of cancers[4, 5, 6, 7].

The adaptive immune system is composed of B and T lymphocytes. These cells differentiate from a common lymphoid progenitor that stems from hematopoietic stem cells in the bone marrow. Pre-B cells continue to mature in the bone marrow, while immature T cells migrate to the thymus to finish developing. Mature B cells express B cell receptors (BCR) that determine their antigen-specificity. Likewise, mature T cells express T cell receptors (TCR). BCRs interact with 3-dimensional epitopes, such as viral or bacterial surface proteins or secreted toxins. On the other hand, TCR specificities are restricted to linear peptide sequences that are presented by major histocompatibility complex (MHC) proteins on the surface of host cells. Activated B cells can secrete their BCRs, interchangeably termed antibodies or immunoglobulin (Ig), to neutralize their target antigens systemically and contribute to humoral immunity. T cells depend on cell-

to-cell (TCR-MHC) contacts and contribute to cell-mediated immunity. The collection of BCRs within an individual is commonly referred to as the BCR or antibody repertoire, and the collection of TCRs is commonly referred to as the TCR repertoire.

Lymphocytes are one of the few types of somatic cells that induce irreversible changes to their genomic DNA on the sequence level. This process, coined V(D)J recombination, occurs during the maturation of B and T cells from their hematopoietic stem cell precursors, resulting in newly formed, naïve B and T cells with unique BCRs and TCRs, respectively. In the case of the BCR heavy chain, individuals have 42-51 distinct functional or open reading frame V genes, 27 D genes, and 6 J genes within their haploid germline genomic DNA[8]. During V(D)J recombination, one V, one D, and one J allele are stitched together (recombined), with the DNA between them being permanently excised. Random nucleotides can be added and/or removed at the junctions. The region surround the D gene is termed the Complementary Determining Region 3 (CDR3) because this short stretch of around 60 base pairs (bp) accounts for a major portion of the diversity and is located at the antigen binding sites. BCRs are made up of two separate polypeptides, the heavy chain and light chain, which are post-translationally joined via covalent di-sulfide bonds. A dimer of two such complexes forms a functional BCR. Both the heavy and light chains recombine independently in a similar manner; however, the light chain lacks the D gene. In humans, the light chain comes in two variants – kappa and lambda – which each have their own collection of V and J genes. Similarly, TCRs are a heterodimer of an alpha and beta chain or gamma and delta chain. As these changes occur on the genomic DNA level, they are passed on to the progeny of the mature lymphocytes, leading to clonal families with identical antigen receptors. Stimulated B cells, but not T cells, can further undergo somatic hypermutation (SHM) – a process

which introduces point mutations into the BCR sequence. Thus, B cell clonal lineages can contain cells with BCR sequences that vary by only one or a few mutations.

V(D)J recombination occurs in the bone marrow for B cells and the thymus for T cells. As their receptors are being rearranged, B and T cells that produce BCRs and TCRs that bind to self-antigens are negatively selected, albeit with less than 100% efficiency. TCRs are further positively selected to weakly associate with MHC molecules. This results in a primary immune repertoire that is imperfectly devoid of self-reactive antibodies and TCRs that are otherwise random in their specificity.

T cells can be broadly separated into two major classes: $CD4^+$ T cells and $CD8^+$ T cells. $CD4^+$ T cells are generally restricted to peptide-MHC class II complexes (pMHC-II). MHC-II are expressed on some antigen presenting cells (APCs) and are loaded with peptides sourced from external proteins. For example, a dendritic cell – one of the key components of the innate immune system – uptakes bacterial proteins through phagocytosis. These proteins are then digested into peptide fragments around 11 – 30 amino acids long and loaded onto MHC-II molecules to form pMHC-II complexes. A $CD4^+$ T cell specific for that bacterial peptide will bind to the pMHC-II complex and initiate downstream signaling to trigger a response. $CD4^+$ T cells carry out a wide range of functions, including helping B cells evolve higher affinity antibodies, activating macrophages, and suppressing inappropriate immune responses (e.g. autoimmunity). They are referred to as "helper T cells" due to their role in assisting the other branches of the immune system.

$CD8^+$ T cells are restricted to pMHC class I (pMHC-I) complexes which are expressed on almost all nucleated cells within the body. Cells that express MHC-I are constantly sampling the proteins within the cell itself. A portion of the proteins being produced within the cell are digested into peptide fragments around 8 – 11 amino acids

long and loaded onto MHC-I molecules. These pMHC-I complexes on the surface of the cells serve as an auditing system for protein production within the cell. As $CD8^+$ T cells that express TCRs that bind to MHC-I molecules loaded with peptides derived from proteins produced in host cells under normal conditions are (mostly) negatively selected, if a $CD8^+$ T cell does bind to a pMHC-I on the surface of a cell, then something must be awry. For example, a virally-infected cell suddenly produces viral proteins. A subset of these proteins are digested, loaded onto MHC-I molecules, and presented on the surface. As these peptides derived from viral proteins are not contained in the set of peptides that $CD8^+$ T cells are negatively selected on, by chance, some $CD8^+$ T cell in the TCR repertoire will express a TCR that binds to this pMHC-I complex. The bound $CD8^+$ T cell indicates that there are inappropriate proteins within the cell, so the $CD8^+$ T cell releases granzymes and perforin to kill off the potentially infected cell. In that sense, $CD8^+$ T cells audit protein product within cells, and the punishment for failing an audit is the death penalty. Effector $CD8^+$ T cells are thus named cytotoxic T lymphocytes (CTLs) for their ability to kill host cells.

1.2 IMMUNOLOGICAL MEMORY AND LYMPHOCYTE DIFFERENTIATION

The defining characteristic of the adaptive immune system is its ability to respond faster and with more fervor upon repeated exposure to an antigen[9]. In essence, the adaptive immune system is able to "remember" past infections, hence the term "immunological memory." This capacity is rooted in the V(D)J recombination process described above and the unique receptors expressed on naive B and T cells. The antibody and TCR repertoires contain hundreds of billions, if not more, of different antibody and TCR sequences. During a primary infection, probabilistically there will be naive B and T cells that are specific for the invading pathogen. However, these antigen-specific naive B

and T cells are rare. As they traffic throughout the lymphatic system, there is a lag phase before the antigen-specific naive cells encounter the antigen, usually in draining lymph nodes, and begin mounting a response. Once activated, the naive cells can proliferate and differentiate into effector phenotypes. Activated B cells can differentiate into plasmablasts and plasma cells and begin to secrete their antigen-specific antibodies. Activated T cells can differentiate into cytokine-producing effector T cells and CTLs. Even after the lag phase, the magnitude of the immune response to an initial infection is relatively low, as measured by the concentration of circulating antibodies specific to the pathogen. It can take weeks for this response to peak.

A subset of the progeny of the activated B and T cells during a primary infection differentiate into memory phenotypes which can persist for decades. Having already engaged antigen during the primary infection, these memory cells are primed and ready for a secondary infection. Upon reactivation, these memory cells differentiate into effector phenotypes in an accelerated manner. There is no lag phase during the secondary infection, and the magnitude of the response is greatly increased over the primary response.

The fate of activated B cells during and after an infection is an area of active research[10]. On the surface level, proliferating B cells enter one of three phenotypes. Short-lived plasma cells differentiate with minimal SHM. These cells are quick to develop and secrete antibodies, and they die off after the infection has been cleared. On the other hand, long-lived plasma cells go through SHM and affinity maturation. These cells traffic to the bone marrow where they can reside for many years while constitutively secreting higher affinity antibodies. Long-lived plasma cells serve as the first line of defense against a secondary infection, as the antibodies already present in the blood can neutralize an infectious agent before it is able to take hold. Lastly, activated B cells can

differentiate into memory B cells. Memory B cells have been observed in individuals 90 years after primary infection[11]. Unlike long-lived plasma cells, memory B cells circulate throughout the lymphatic system and peripheral blood until the come back into contact with their target antigens. Then, they can quickly differentiate into plasma cells to secrete high concentrations of antibodies.

While they are part of the immune network which is intricately connected, the primary role of most B cell phenotypes is to secrete antibodies to fight infections. T cells, on the other hand, have a much more nuanced and diverse role. The differences between $CD4^+$ and $CD8^+$ T cells highlighted above only begin to scratch the surface of this heterogeneity. This is reflected in the large number of T cell memory subsets with varying functionalities. T cells are often categorized by their migratory patterns and cytokine expression profiles[12]. Central memory T cells (T_{CM}) express CCR7 and CD62L and home to secondary lymphoid tissues. Effector memory T cells (T_{EM}) lack CCR7 and CD62L expression and tend to circulate through the periphery. Tissue resident memory T cells (T_{RM}) reside directly in tissues likely to be exposed to pathogens, such as the lungs, skin, and gut mucosa. T_{CM} retain a naive-like stemness in their ability to proliferate rapidly when activated, while T_{EM} are skewed more towards effector function and are more proficient at releasing cytokines.

$CD4^+$ T cells are particularly heterogeneous[13]. Depending on the cytokine context in which naive $CD4^+$ T cells are activated, they can skew their cytokine expression profiles. IFN-γ and IL-12 lead to upregulation of T-bet which is the master transcription factor of the T_H1 phenotype. T_H1 cells predominantly produce IFN-γ and IL-2 and work to activate macrophages and $CD8^+$ T cells. IL-4 and IL-2 promote GATA3 upregulation which leads to a T_H2 phenotype. T_H2 cells produce IL-4 which drives B cell proliferation and encourages isotype switching. TGF-β and IL-6 lead to RORγt upregulation which

governs the T_H17 response. T_H17 cells produce IL-17 which has broad, proinflammatory functionality. An important subset of $CD4^+$ T cells, regulatory T cells (T_{regs}) express the transcription factor Foxp3 and produce IL-10 and TGF-β to abrogate inappropriate immune responses. T_{regs} are largely responsible for limiting autoimmunity.

Finally, follicular helper T cells (T_{FH}) are specialized $CD4^+$ T cells that provide help to germinal center B cells[14]. Bcl-6 serves as the master regulator transcription factor for T_{FH} cells, inhibiting the expression of T-bet, GATA3, and RORγt of other differentiation pathways. T_{FH} express the chemokine receptor CXCR5 which leads to migration towards CXCL13 produced in the B cell follicles within lymph nodes. Thus, T_{FH} colocalize with activated B cells within the B cell follicle and spur the formation of germinal centers. T_{FH} express several costimulatory molecules, such as CD40L, SAP, ICOS, OX40, and PD-1 which interact with CD40, SLAM, ICOSL, OX40L, and PD-L1/2 on activated B cells, respectively, to aid in germinal center B cell proliferation, maturation, class switching, and survival. T_{FH} produce IL-21 and IL-4 which further drive germinal center B cell development into plasma cells and memory B cells. It is becoming increasingly clear that even within these specialized $CD4^+$ T cell subsets there is much diversity.

1.3 IMMUNE SYSTEM DEVELOPMENT WITH AGE

The immune system beings to develop in the fetus and continues well into adolescence before settling into adult-like properties after puberty. Understanding the immune system's evolution with age and capacity to fight pathogens is critical in our quest to lower infant mortality and improve health standards, as infections are one of the leading causes of death in young children worldwide. The various cogs of the immune system develop with different kinetics. As these cogs must work together in unison to

provide protection, minor changes in one aspect may have wide-ranging effects on the immune system as a whole. Drug interventions that are suitable for adults may not be optimal for infants depending on the developmental stage of their immune systems. Better knowledge of these stages of development could lead to more treatment options tailored to the specific stages which in turn could lead to better clinical outcomes. For example, the currently most advanced malaria vaccine, RTS,S/AS01, has been observed to be protective in only 18% of cases in young infants, compared to 28% in children[15], leaving much room for improvement. Discovering the exact source of this 10% discrepancy is the first step to overcoming it. Studying the immune system at different life stages has the potential to unlock new therapies that compensate for inherent deficiencies.

These inherent deficiencies are widespread throughout both the innate and adaptive immune cells[16]. With respect to the innate immune system, neutrophils arise *in utero* and reach adult-like counts within days of birth, but they are functionally impaired. Neutrophils have reduced phagocytic capabilities at birth, but they are quickly restored within 3 days[17]. However, neonatal neutrophils are also deficient in CD11b expression and exhibit reduced rolling and adhesive functions until 11-12 months of age[18, 19]. Despite similar cell counts, preterm neonatal plasmacytoid dendritic cells secrete significantly less IFN-α in response to TLR9 engagement, likely contributing to their exacerbated symptoms during viral infections such as RSV, HSV, and CMV[20].

On the adaptive side, neonatal $CD4^+$ T cells are predisposed towards T_{reg} and T_H2 phenotypes[16]. This leads to more tolerance and a weaker cell-mediated response to foreign pathogens. Neonatal B cells express lower levels of B7-1/2 and CD40, while neonatal T cells express lower levels of CD40L upon activation, likely contributing to impaired humoral immunity[21]. Infant B cells are particularly unresponsive to T-cell-

independent antigens. This response is largely driven by splenic marginal-zone B cells which do not appear in significant numbers until 1-2 years old[22]. Germinal center formation may also be impaired in young infants, diminishing the B cell response to T-cell-dependent antigens as well[23]. Overall, many facets of the infant immune system are still works in progress at birth, but the implications of these deviations from adult-like features are not fully elucidated.

1.4 IMMUNE REPERTOIRE SEQUENCING

The advent of high-throughput sequencing technologies has opened the door to a new dimension of research in biology[24]. Cheaper and cheaper access to technology capable of generating millions to billions of sequencing reads has redefined the term "Big Data," and the implications have only just begun to be explored. Within the realm of immunology, high-throughput sequencing platforms like the Illumina MiSeq have kickstarted the field of Immune Repertoire Sequencing (IR-Seq). IR-Seq experiments seek to characterize the antibody and/or TCR repertoire via high-throughput sequencing of the multitude of diverse antigen receptors generated through V(D)J recombination.

Early antibody sequencing experiments were limited to only dozens of sequences[25]. Given the sheer number of possible V(D)J rearrangements, junction insertions/deletions and SHM notwithstanding, characterizing the antibody repertoire to any meaningful depth was impossible. However, as technology improved and this restriction was lifted, another issue arose: sequencing errors. As discussed in **section 1.1**, clonal expansion and SHM can lead to related antibody sequences that vary by only one or a few mutations. The reported error rate for Illumina MiSeq is about 0.5% errors per bp[26]. For an antibody sequencing read over 200bp long, this equates to 1 error per

sequencing read on average. Thus, distinguishing between genuine mutations and sequencing errors is a nontrivial task.

IR-seq experiments tend to follow a similar basic framework. Most begin with reverse transcription of antibody or TCR mRNA, though some utilize genomic DNA. The resulting cDNA (or genomic DNA) must then be amplified to generate enough material for sequencing. This is achieved through many cycles of PCR, which can lead to amplification bias and PCR errors that further confound the inherent sequencing errors. Adaptors necessary for integration with the sequencing platform are added through PCR with overhanging primers to generate the final sequencing library.

Molecular identifiers (MIDs) have been used to mitigate PCR and sequencing errors to overcome this pitfall of early next-generation antibody repertoire sequencing experiments[27, 28, 29, 30]. MIDs are short stretches of randomized DNA sequences that are tagged to individual cDNA molecules during reverse transcription and/or second strand synthesis. This can be achieved by fusing the randomized MIDs to the reverse transcription primer followed by a common sequencing adaptor sequence. The sequencing adaptor is then used for downstream PCR amplifications such that the randomized MID is conserved for each cDNA molecule that originated from the same mRNA transcript. Sequencing reads are then grouped according to the MID, and consensus sequences are built by taking the average nucleotide at each position. This essentially averages out PCR and sequencing errors that plagued early high-throughput IR-Seq experiments. For example, consider a single antibody mRNA transcript represented by 5 sequencing reads. Each raw read is inflicted by a different sequencing error. Without MIDs, these 5 sequencing reads would appear to be 5 distinct, clonally related antibody sequences. However, each of the 5 reads has the same MID, so they can be grouped together. For each of the 5 sequencing errors, the other 4 reads have the

correct nucleotide, so the consensus sequence constructed by averaging the sequencing reads will have the correct full-length antibody sequence. MIDs cannot compensate for errors during reverse transcription, but they provide the accuracy necessary for high quality repertoire characterization.

1.5 REFERENCES

1. Barouch DH*, et al.* Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* **503**, 224-228 (2013).

2. Nguyen AW*, et al.* A cocktail of humanized anti-pertussis toxin antibodies limits disease in murine and baboon models of whooping cough. *Science translational medicine* **7**, 316ra195 (2015).

3. Corti D*, et al.* Protective monotherapy against lethal Ebola virus infection by a potently neutralizing antibody. *Science* **351**, 1339-1342 (2016).

4. Topalian SL, Drake CG, Pardoll DM. Immune checkpoint blockade: a common denominator approach to cancer therapy. *Cancer cell* **27**, 450-461 (2015).

5. Davila ML*, et al.* Efficacy and toxicity management of 19-28z CAR T cell therapy in B cell acute lymphoblastic leukemia. *Science translational medicine* **6**, 224ra225 (2014).

6. Hamid O*, et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *The New England journal of medicine* **369**, 134-144 (2013).

7. Weber JS*, et al.* Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial. *The Lancet Oncology* **16**, 375-384 (2015).

8. Watson CT*, et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *American journal of human genetics* **92**, 530-546 (2013).

9. Ahmed R, Gray D. Immunological memory and protective immunity: understanding their relation. *Science* **272**, 54-60 (1996).

10. Kurosaki T, Kometani K, Ise W. Memory B cells. *Nature reviews Immunology* **15**, 149-159 (2015).

11. Yu X*, et al.* Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* **455**, 532-536 (2008).

12. Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T cell subsets, migration patterns, and tissue residence. *Annual review of immunology* **31**, 137-161 (2013).

13. Zhu J, Yamane H, Paul WE. Differentiation of effector CD4 T cell populations (*). *Annual review of immunology* **28**, 445-489 (2010).

14. Crotty S. Follicular helper CD4 T cells (TFH). *Annual review of immunology* **29**, 621-663 (2011).

15. Efficacy and safety of the RTS,S/AS01 malaria vaccine during 18 months after vaccination: a phase 3 randomized, controlled trial in children and young infants at 11 African sites. *PLoS medicine* **11**, e1001685 (2014).

16. Simon AK, Hollander GA, McMichael A. Evolution of the immune system in humans from infancy to old age. *Proceedings Biological sciences* **282**, 20143085 (2015).

17. Filias A, Theodorou GL, Mouzopoulou S, Varvarigou AA, Mantagos S, Karakantza M. Phagocytic ability of neutrophils and monocytes in neonates. *BMC pediatrics* **11**, 29 (2011).

18. Storm SW, Mariscalco MM, Tosi MF. Postnatal maturation of total cell content and up-regulated surface expression of Mac-1 (CD11b/CD18) in polymorphonuclear leukocytes of human infants. *Journal of leukocyte biology* **84**, 477-479 (2008).

19. Nussbaum C*, et al.* Neutrophil and endothelial adhesive function during human fetal ontogeny. *Journal of leukocyte biology* **93**, 175-184 (2013).

20. Schuller SS*, et al.* Preterm neonates display altered plasmacytoid dendritic cell function and morphology. *Journal of leukocyte biology* **93**, 781-788 (2013).

21. Kaur K, Chowdhury S, Greenspan NS, Schreiber JR. Decreased expression of tumor necrosis factor family receptors involved in humoral immune responses in preterm neonates. *Blood* **110**, 2948-2954 (2007).

22. Adkins B, Leclerc C, Marshall-Clarke S. Neonatal adaptive immunity comes of age. *Nature reviews Immunology* **4**, 553-564 (2004).

23. Timens W, Rozeboom T, Poppema S. Fetal and neonatal development of human spleen: an immunohistological study. *Immunology* **60**, 603-609 (1987).

24. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **32**, 158-168 (2014).

25. Ridings J, Dinan L, Williams R, Roberton D, Zola H. Somatic mutation of immunoglobulin V(H)6 genes in human infants. *Clinical and experimental immunology* **114**, 33-39 (1998).

26. Loman NJ*, et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology* **30**, 434-439 (2012).

27. Shugay M*, et al.* Towards error-free profiling of immune repertoires. *Nature methods*, (2014).

28. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 13463-13468 (2013).

29. Vander Heiden JA*, et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, (2014).

30. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* **2**, e1501371 (2016).

Chapter 2: Molecular Identifier Clustering-based Immune Repertoire Sequencing[1]

2.1 INTRODUCTION

V(D)J recombination creates hundreds of billions of antibodies and T cell receptors that collectively serve as the immune repertoire to protect the host from pathogens. Somatic hypermutation (SHM) further diversifies the antibody repertoire, which makes it impossible to quantify this diversity with nucleotide resolution until the development of high-throughput sequencing-based Immune Repertoire Sequencing (IR-Seq)[1, 2, 3, 4]. Although we and others have developed methods to control for artifacts from high amplification bias and sequencing error rates through data analysis[3, 5, 6, 7, 8, 9], obtaining accurate sequencing information has now been made possible by the use of molecular identifiers (MID)[10, 11, 12, 13]. MIDs serve as barcodes to track genes of interest through amplification and sequencing. They are short stretches of nucleotide sequence tags composed of randomized nucleotides that are usually tagged to cDNA during reverse transcription to identify sequencing reads that originated from the same mRNA transcript.

There still lacks a general framework with respect to both experimental design and bioinformatics analysis on how to use MIDs, for example: what is the minimum cell input amount, how to efficiently use MIDs to tag each transcript, how to group reads to generate consensus sequences, and what quality metrics one can use to check their IR-Seq methods. Answers to these questions are important for overall experiment design, repertoire diversity estimates, and controlling the accuracy of the sequence information

[1]Wendel, *et al.* Accurate Immune Repertoire Sequencing Reveals Malaria Infection Driven Antibody Lineage Diversification in Young Children. *Nature Communications,* accepted. B.S.W. performed naive B cell isolation and data analysis; C.H. helped with sequencing data analysis; M.Q. developed the sequencing protocol using sorted naive B cells; D.W. developed the pipeline; P.R. provided computation resources and helped with analysis; N.J. conceived the idea, designed the study, and directed data analysis; B.S.W. and N.J. wrote the paper with contributions from all co-authors.

obtained. Despite recent advancements, the large amount of input RNA required and low diversity coverage make it challenging to analyze small numbers of cells, such as memory B cells from dissected tissues or blood draws from young children, using IR-Seq because these samples require many PCR cycles to generate enough material to make sequencing libraries, thus exacerbating PCR bias and errors.

Here, we report the development of MID clustering-based IR-Seq (MIDCIRS) that further separates different RNA molecules tagged with the same MID. Using naive B cells, we demonstrate that MIDCIRS has a high coverage of the diversity estimate, or different types of antibody sequences, that is consistent with the input cell number and a large dynamic range of three orders of magnitude compared to other MID-based immune repertoire sequencing methods[10, 11]. Given the wide use of IR-Seq in basic research as well as clinical settings, we believe the method outlined here will serve as an important guideline for future IR-Seq experimental designs.

2.2 RESULTS

2.2.1 Overview of the MIDCIRS method

MIDs have been used to track individual RNA molecules through PCR and sequencing in IR-Seq to reduce error rate[10, 11, 12, 13]. They can be designed with sufficient length, and thus diversity, to tag each individual molecule uniquely. However, this requires knowledge of the total number of RNA molecules in the sample, which requires a significant amount of time and effort to measure for each sample before library preparation. Longer MIDs are likely to decrease the reverse transcription efficiency[14, 15]. Therefore, we fixed the MID length at 12 random nucleotides and developed a generalized approach to identify each individual transcript using a sequence-similarity-

based clustering method to separate a group of sequencing reads with the same MID into sub-groups (**Figure 2.1**).

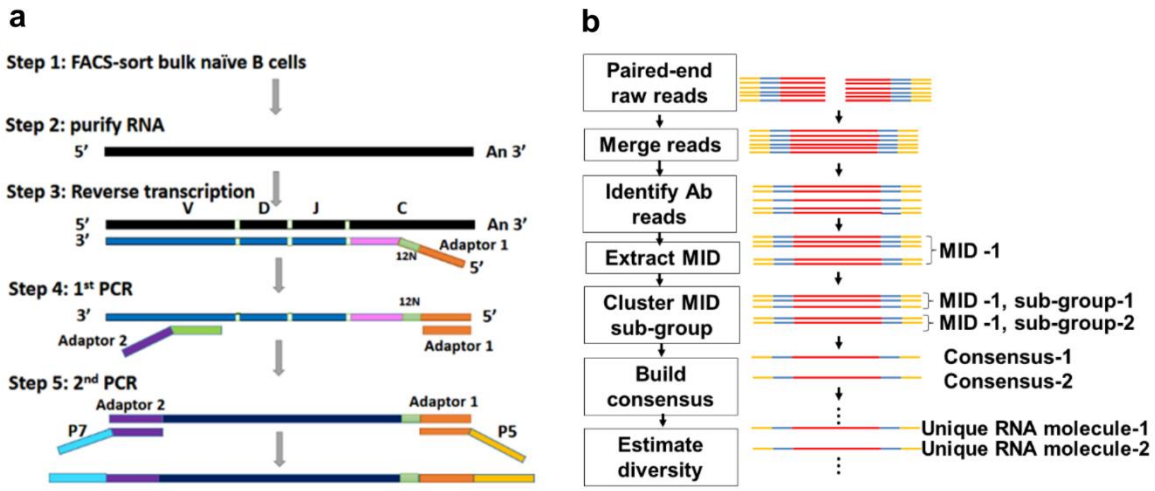

Figure 2.1: MIDCIRS overview. (**a**) Schematic overview of tagging single antibody transcripts with MIDs. (**b**) Schematic overview of the informatics pipeline of MIDCIRS which includes merging paired-end reads, performing clustering to generate MID sub-groups, and building consensus sequences.

In brief, we tag MIDs to cDNA during reverse transcription by fusing the 12N MIDs and a partial sequencing adaptor to antibody heavy chain constant region-specific primers. We then use multiplexed primers complementary to the 5' end of the antibody V gene alleles fused to another partial sequencing adaptor to amplify the cDNA in the 1st PCR step. We perform a 2nd PCR step to ligate the full adaptor sequences before gel purifying, quantifying, and sequencing the library on the Illumina MiSeq 2x250 platform. Primers used for library generation can be found in **Table A.1**. The resulting paired-end sequencing reads are merged, and antibody reads are identified. Antibody reads are then grouped according to MID, and then each MID is split into sub-groups by clustering on sequence similarity. This clustering step separates distinct RNA molecules that, by

chance, were labeled with the same MID during reverse transcription. Consensus sequences are then built by taking the average nucleotide at each position within a sub-group, weighted by the quality score. Each consensus sequence represents an mRNA molecule, and identical consensus sequences are merged into unique consensus sequences, or unique RNA molecules.

2.2.2 Sub-group clustering threshold determination

The clustering threshold for sub-group generation must be lenient enough to group reads that differ due to PCR and sequencing errors into the same MID sub-group but stringent enough to exclude reads that are derived from different antibody sequences. We used RNA controls with known sequences to ensure the Levenshtein distance threshold (number of substitutions, insertions, or deletions required to match 2 strings) was lax enough to group 99% of the control sequences together. For both control templates, this was achieved at a Levenshstein distance of 23 for a 150bp sequence, so the clustering threshold was set to 15% of the read length (**Figure 2.2a**). To test whether this threshold was strict enough to prevent distinct antibody sequences from grouping together, we used a library of 10^6 FACS-sorted naive B cells and plotted the percentage of MID groups with multiple sub-groups at varying clustering thresholds (**Figure 2.2b**). At 0% clustering threshold, a single PCR or sequencing error would result in sub-group formation, so the percentage of MID groups containing multiple sub-groups is at a maximum. By 60% clustering threshold, even antibody sequences utilizing different V, D, and J genes can be grouped together, so the number of MIDs with multiple sub-groups approaches 0, and the clustering step loses distinguishing power. The 15% clustering threshold occurs just before the steep drop in the sigmoidal curve, indicating that it is unlikely that many distinct antibody sequences were merged.

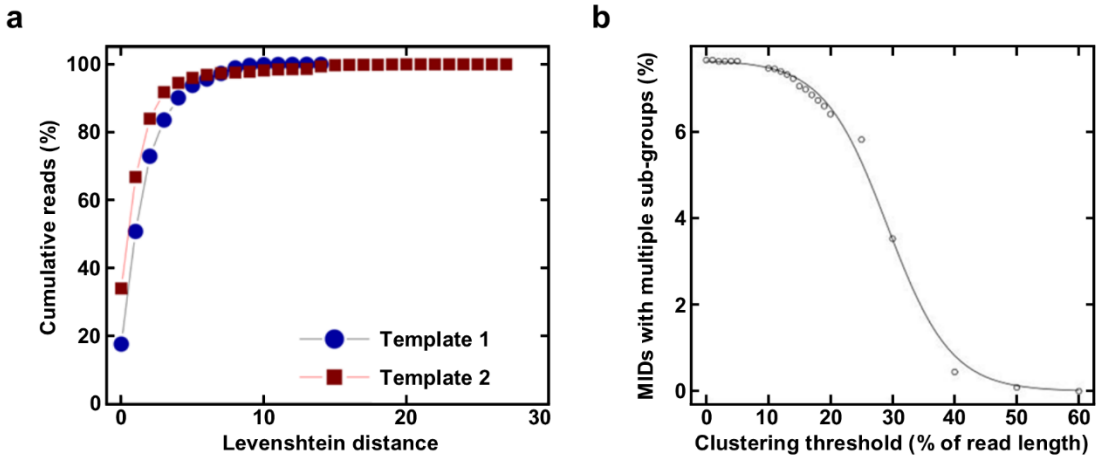

Figure 2.2: Sub-group clustering threshold calibration. (**a**) Cumulative percentage of reads as a function of the Levenshtein distance between the RNA control templates and sequencing reads. The lengths of the control templates and reads were 150bp. (**b**) Percentage of MIDs with multiple sub-groups as a function of the clustering threshold.

2.2.3 MIDCIRS yields high accuracy and coverage down to 1000 cells

We used FACS-sorted naive B cells with varying numbers (10^3 to 10^6) to test the dynamic range of MIDCIRS. Overall, 95% of the sequencing reads were successfully merged, and of those 98% were assigned as antibody sequences (**Table 2.1**). These percentages were generally higher at larger cell inputs, but the absolute number of erroneous reads at low cell counts was miniscule in terms of cost and computational resources.

Our digital PCR results (not shown) give a lower end of 10 antibody RNA copies per naive B cell. With 30% RNA input, for 10^6 naive B cells, this yields $3x10^6$ antibody RNA molecules being tagged by 4^{12} MIDs. Poisson statistics predicts that 8.5% of the MIDs would tag 2 or more different RNA molecules. This model closely matches our experimental results, with 7.1% of the MIDs in the 10^6 naive B cell library containing

multiple sub-groups. Without the sub-group clustering step of MIDCIRS, these 7.1% of MIDs would either be discarded or lead to incorrect consensus sequence formation. As expected, the number of MIDs containing multiple sub-groups decreases with decreasing input material, and the 10^3 naive B cell library only has a single MID with multiple sub-groups.

Cells (#)	Raw reads (#)	Merged reads (#)	Antibody reads (#)	Useful MIDs[a] (#)	Reads in useful MIDs (%)	Multi-sub-group MIDs[b] (#)
1,000	46,320	22,742	9,201	797	94.30	1
2,000	44,846	18,602	17,421	2,176	93.29	2
10,000	228,711	99,370	62,242	7,102	94.73	9
20,000	293,279	196,570	184,754	23,991	93.27	49
100,000	1,153,763	1,074,771	1,048,523	165,663	92.63	1,137
200,000	2,191,738	2,107,762	2,059,944	404,225	91.41	7,239
1,000,000	7,494,809	7,342,163	7,258,253	1,516,098	86.44	108,172

Table 2.1: Sequencing reads statistics for naive B cell libraries. [a]A useful MID has more than two reads. If there are only two reads in an MID, they are discarded unless they are identical. [b]The number of MIDs containing sequences derived from 2 or more different antibody sequences.

An important but often overlooked parameter for IR-Seq experiments is the diversity coverage, or the percentage of the maximum number of unique sequences discovered. While a lack in diversity coverage can be overcome by increasing the cell input to obtain enough sequences for comprehensive repertoire analysis, this is not always possible with precious samples and rare cell types, for example memory B cells sorted from infants from whom only 4mL of blood can be ethically drawn, and memory B cells are rare (see **Chapter 3**). To estimate the diversity coverage of MIDCIRS, we plotted the number of unique RNA molecules discovered against the cell input number (**Figure 2.3a**). The results displayed a strong correlation with cell numbers at 83%

coverage (**Figure 2.3a**, slope). Previous studies have shown that about 80% of naive B cells express distinct heavy chain genes[16, 17], thus our method achieves a comprehensive diversity coverage that is much higher than other MID-based antibody repertoire sequencing techniques[10, 11, 12, 13] .

Next, we examined the error rate with or without using MIDCIRS. Because the diversity among hundreds of millions of antigen receptors lies in a short stretch of DNA about 60 nucleotides, often two distinct sequences differ by only a few nucleotides. In addition, somatic hypermuation, a process that further diversifies antibody sequences, has a mutation rate that is comparable to the error rate of the next-generation sequencers. This makes estimating the total antigen receptor diversity and tracing the mutational evolution of antibody sequences difficult. We calculated the raw read error rate without MIDCIRS by comparing the individual reads within a sub-group to the consensus sequence and reached a similar rate as previously reported for Illumina sequencing, about 0.5%[18] (**Figure 2.3b**, top panel). To calculate the improved error rate using MIDCIRS, we split the total reads into two groups, performed clustering separately, and then compared the consensus of overlapping sub-groups from these two sub-samples. The resulted error rate was 130-fold lower than the raw error rate (**Figure 2.3b**, bottom panel). In addition, while the raw error rate fluctuated between the 3 runs (**Figure 2.3b**, top panel), the improved error rate after using MIDs remained constant (**Figure 2.3b**, bottom panel). Taken together, these results demonstrate that MIDCIRS can cover a large dynamic range on cell input with remarkable coverage while maintaining a low error rate.

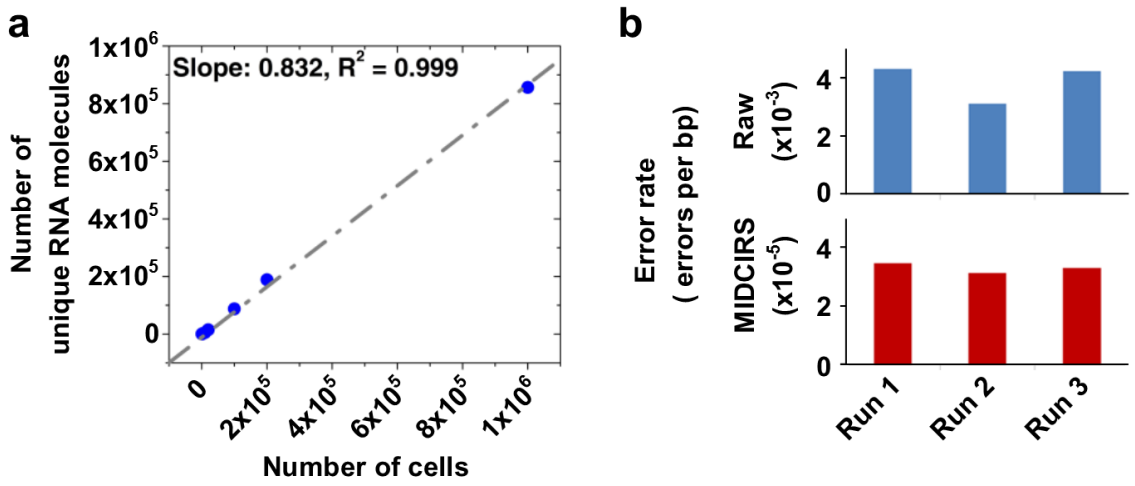

Figure 2.3: MIDCIRS yields high diversity coverage with low error rate. (**a**) Correlation between number of cells and number of unique RNA molecules after using MIDCIRS. RNA from as few as 1,000 to as many as 1,000,000 naive B cells was used as input material in generating the amplicon libraries. Slope indicates the estimated diversity coverage. (**b**) Comparison between raw error rate and improved error rate after using MIDCIRS. Raw reads error rates (top, blue) and MIDCIRS consensus error rates (bottom, red) for 3 Miseq runs, calculated as described in **Methods 2.4.5**.

2.2.4 MIDCIRS is robust and mitigates artificial diversity

Sequencing depth is another important factor to consider when designing an IR-Seq experiment. In order to utilize MIDs to reduce the error rate, IR-Seq libraries must be sequenced deep enough to ensure that most MIDs have at least 2 reads associated with them. Consensus building cannot be performed on a single read, so MIDs with only a single read must be discarded, as they are plagued by the inherently high raw reads error rate shown in **Figure 2.3b**. Here, we sequenced each library to a depth of 5 times the cell input number and found that 92% of the reads had useful MIDs (**Table 2.1**). While suitable for these FACS-sorted naive B cell libraries, the optimal sequencing depth varies depending on the sample type and number of antibody transcripts. More antibody transcripts will require more reads in order to saturate.

As a generalized approach to IR-Seq experimental quality control and sequencing depth check, we performed rarefaction analysis by subsampling the sequencing reads to different amounts and then computing the diversity to test the effect of sequencing depth and error rate on MIDCIRS. On average, the rarefaction curves reach a plateau at a sequencing depth of around three times the cell number using MIDCIRS, suggesting that sequencing more will not discover further diversity (**Figure 2.4a**). In contrast, without using MIDCIRS, the number of unique sequences continues to increase well beyond the number of cells for all samples (**Figure 2.4b**). This artificial diversity from oversequencing is a result of PCR or sequencing error-inflicted sequences being discovered which were hidden at lower sequencing depth due to their relative rarity. These rare errors are averaged out during consensus building, so MIDCIRS effectively mitigates artificial diversity. The optimal sequencing depth is likely to change depending on sample composition, so rarefaction analysis should be performed on every newly generated IR-Seq library.

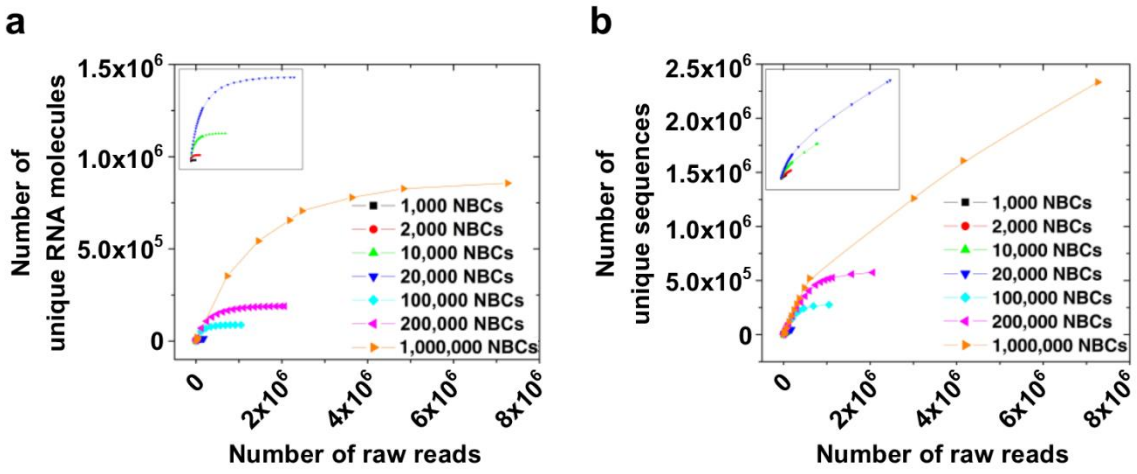

Figure 2.4: Rarefaction analysis for optimal sequencing depth. Rarefaction analysis for each library with (**a**) and without (**b**) using MIDCIRS. Inset displayed zoomed-in view near the origin for low cell count libraries. NBCs, naive B cells.

2.3 DISCUSSION

IR-Seq has been very useful to study the immune response in both research and clinical settings[19, 20]. However, the first generation of IR-Seq suffered from high PCR and sequencing errors, making it hard to distinguish true somatic hypermutations from errors or true CDR3 diversity from artifacts. In addition, the IR-Seq field lacks a general framework that one can follow when designing and validating any new IR-Seq methods. In this study, we described a new IR-Seq method, MIDCIRS, which separates different RNA molecules that might have been labeled with the same MIDs into different sub-groups. This method is robust, accurate, and suitable for analyzing small amount of cell input, while correcting undercounted diversity.

Ensuring that each MID only labels one specific RNA molecule requires prior knowledge of the total antibody RNA transcript number in the sample, which takes time and resources to measure and requires using a portion of the RNA material which could

otherwise be used for other experiments to maximize the data output from precious samples. Here, we demonstrated that, by using clustering based on MID groups, we can identify new diversities, and this corrected diversity matches cell input numbers very well. We showed that when input cell numbers are low, most of the MIDs contain only one type of heavy chain transcript. However, as the cell number increases, there are more MIDs containing more than one type of antibody heavy chain transcript. For 10^6 naïve B cells, 108,172 out of 1,516,098 useful MIDs (7.1%) tagged 2 or more unique RNA molecules. Without sub-group clustering, the diversity of this library would be significantly undercounted. This number is likely to increase if more cells are used as an input material or B cells containing high antibody transcript counts constitute a significant portion of the input material, such as plasmablasts in PBMC samples after vaccination. Therefore, directly building consensus using MIDs is likely to underestimate the total diversity. Using our MIDCIRS method, we analyzed data from a recent publication[11] where the same number of random nucleotides (12N) was used. Using the same criteria defining unique CDR3 sequences, our MIDCIRS analysis showed that about 5% of the diversity was undercounted because sequencing reads from multiple distinct RNA molecules that were tagged with the same MID were counted as one type of RNA by Shugay et al[11], highlighting the necessity of the clustering method we developed in MIDCIRS.

In summary, we presented a new method of utilizing MIDs to significantly improve the accuracy and robustness of the IR-Seq and approaches that one can incorporate into their experiment designs to QC the method and analyses. We believe that this framework will enhance the application of IR-Seq in basic and clinical research.

2.4 METHODS

2.4.1 Sample collection and isolation

Human PBMCs were purified from de-identified blood bank donor samples using conventional Ficoll density gradient centrifugation. This protocol was approved by the Institutional Review Board of the University of Texas at Austin as non-human subject research. Naive B cells were FACS-sorted at varying cell counts (10^3 to 10^6) based on the phenotype of $CD3^-CD19^+CD20^+CD27^-CD38^{low}$. The following antibody clones were obtained from Biolegend and used as 1:25 dilution: Alexa Fluor 488 OKT3 (CD3), Pacific Blue 2H7 (CD20), Brilliant Violet 605 O323 (CD27), and APC-Cy7 HIT2 (CD38). Flow cytometry and cell sorting were performed on BD FACS Aria II.

2.4.2 Bulk antibody sequencing library generation and sequencing

MIDs were added during the reverse transcription step through the use of fusion primers, which contain the partial Illumina P5 sequencing adaptor followed by twelve random nucleotides and primers to the constant region of five antibody isotypes. We fused eleven forward primers that were previously designed[6] to partial Illumina P7 adaptor. Full Illumina adaptors were added during the second PCR step along with library indexes. Total RNA was purified using All Prep DNA/RNA kit (Qiagen) following the manufacturer's protocol. cDNA synthesis was done using Superscript III (Life Technologies) with 30% of the RNA as input. After free primer removal by Exonuclease I digestion (New England Biolabs) according to the manufacturer's protocol, Takara Ex Taq HS polymerase (clone Tech) was used for both PCR reactions. The first PCR was performed with the following program: initial denature at 95°C for 3 minutes, followed by 20 cycles of 95°C for 30 seconds, 57°C for 30 seconds, and finally 72°C for 2 minutes with a 4°C hold. All cDNA was used for first PCR and about 1% of 1st PCR product was

used for second PCR. The second PCR was performed with the following program: initial denature at 95°C for 3 minutes, followed by 10 cycles of 95°C for 30 seconds, 57°C for 30 seconds, and finally 72°C for 2 minutes with a 4°C hold. The second PCR products were gel purified and quantified by qPCR Library Quantification Kit (KAPA biosystems, catalog number KK4824). 0.009 pmol of the QCed second PCR product were sequenced on Illumina Mi-seq with paired-end 250bp read mode (sequencing kit catalog number: MiSeq Reagent Kit v2 (500cycle) MS-102-2003). The list of primers for RT and PCR can be found in **Table A.1**. Libraries were sequenced multiple times until saturated based on rarefaction analysis in **Figure 2.4**. Reads from all runs were combined and analyzed.

2.4.3 Preliminary read processing

Raw reads from Illumina MiSeq PE250 were first cleaned up following steps outlines in **Figure 2.1b**. Only reads that exactly matched the corresponding library indices were included for further processing. The end of each raw read was trimmed such that all bases had a quality score of 25 or higher. Reads 1 and 2 were merged using the SeqPrep tool (https://github.comjstjohn/SeqPrep). The merged reads were filtered with specific V-gene and constant region primers to determine antibody reads. The primers were then truncated from the reads. The retained reads were further truncated to 320bp. Read numbers after each filter are listed in **Table 2.1**.

2.4.4 MID sub-group generation

Raw reads were split into MID groups according to their 12 nucleotide barcodes. For each MID group, quality threshold clustering was used to cluster similar reads. This process groups reads derived from a common template RNA molecule together while separating reads derived from distinct RNA molecules. A Levenshtein distance of 15% of the read length was used as the threshold. This was calibrated using RNA controls with

known sequences (**Figure 2.2**). For each sub-group, a consensus sequence was built based on the average nucleotide at each position, weighted by the quality score. In the case that there were only two reads in an MID sub-group, we only considered them useful reads if both were identical. Each MID sub-group is equivalent to an RNA molecule. Next, we merged all of the identical consensus to form unique consensus sequences, or unique RNA molecules, which were used to estimate the diversity and assess the sequencing depth in rarefaction analysis (**Figure 2.4**). Scripts for this section can be downloaded at https://github.com/utjianglab/MIDCIRS.

2.4.5 Error rate calculation

The difference between the consensus sequence for an RNA molecule and the raw reads associated with it represent the errors generated in either PCR or sequencing. The error rate can be calculated using the following formula (2.1):

$$ErrorRate(Raw) = \frac{\sum_{i=1}^{N_I} Diff(i,I)}{N_I \times L} \quad (2.1)$$

where *Diff(i,I)* is the Levenshtein distance between read *i* and the consensus sequence in MID sub-group *I*, N_I is the number of reads in MID sub-group *I*, and *L* is the read length.

In order to estimate the improved error rate using MIDCIRS, we equally divided the raw reads from one library into two datasets. The same MID sub-group generating process was performed on both datasets. By comparing the differences between the consensus sequences with identical MID between these two datasets, we can calculate the improved error rate for using MID sub-groups as:

$$ErrorRate(MID) = \frac{\sum_{I,J} Diff(I,J) \times N_I}{\sum_I N_I \times L} \quad (2.2)$$

where *Diff(I,J)* is the Levenshtein distance between the consensus *I* and consensus *J* which have the identical MID, N_I is the number of reads in MID sub-group *I*, and *L* is the read length.

2.5 REFERENCES

1. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **32**, 158-168 (2014).

2. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Current opinion in immunology* **25**, 646-652 (2013).

3. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807-810 (2009).

4. Tipton CM*, et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nature immunology* **16**, 755-765 (2015).

5. Jiang N, Weinstein JA, Penland L, White RA, 3rd, Fisher DS, Quake SR. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5348-5353 (2011).

6. Jiang N*, et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine* **5**, 171ra119 (2013).

7. Bolotin DA*, et al.* Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *European journal of immunology* **42**, 3073-3083 (2012).

8. Michaeli M, Noga H, Tabibian-Keissar H, Barshack I, Mehr R. Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front Immunol* **3**, 386 (2012).

9. Zhu J*, et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains.

Proceedings of the National Academy of Sciences of the United States of America **110**, 6470-6475 (2013).

10. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 13463-13468 (2013).

11. Shugay M*, et al.* Towards error-free profiling of immune repertoires. *Nature methods*, (2014).

12. Vander Heiden JA*, et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, (2014).

13. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* **2**, e1501371 (2016).

14. Shiao YH. A new reverse transcription-polymerase chain reaction method for accurate quantification. *BMC Biotechnology* **3**, 22 (2003).

15. Zajac P, Islam S, Hochgerner H, Lonnerberg P, Linnarsson S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS One* **8**, e85270 (2013).

16. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116**, 1070-1078 (2010).

17. DeKosky BJ*, et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature biotechnology* **31**, 166-169 (2013).

18. Loman NJ*, et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30**, 434-439 (2012).

19. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* **32**, 158-168 (2014).

20. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr Opin Immunol* **25**, 646-652 (2013).

Chapter 3: Accurate Immune Repertoire Sequencing reveals malaria infection-driven antibody lineage diversification in young children[2]

3.1 INTRODUCTION

The immune system, similar to the vessel that houses it, develops with age[1]. Infants are born fragile and vulnerable. All of the parts are present – eyes, ears, B cells, T cells – but they are not fully matured. While the maturation of the physical body is obvious, the maturation of the immune system is more subtle. Some components of the innate immune system, such as neutrophils and pulmonary macrophages, reach adult-like levels within days of birth but exhibit reduced functionality[2, 3]. Aspects of the adaptive immune system are also impaired early in life. B cell responses to T cell-independent polysaccharide antigens are notably impaired until 1-2 years of age, and antibody responses to T cell-dependent antigens are also diminished in infants[4]. During this time, infants routinely receive vaccines and are exposed to environmental antigens. Umbilical cord blood provides a noninvasive biospecimen to sample the newborn immune system; however, the post-birth development of the immune system is not well characterized due in large part to the limited sample size availability. In particular, there lacks a comprehensive analysis of the antibody repertoire changes that occur during a natural infection.

---

[2]Wendel, *et al.* Accurate Immune Repertoire Sequencing Reveals Malaria Infection Driven Antibody Lineage Diversification in Young Children. *Nature Communications,* accepted. B.S.W. performed all malaria related experiment and data analysis; C.H. performed BASELINe and other sequencing data analysis; M.Q. developed the sequencing protocol using sorted naive B cells; D.W. helped with sequence analysis; S.M.H. helped with library construction; K.Y.M. helped with sequencing; J.X. helped with lineage visualization, E.W.L, P.D.C., and S.K.P. selected malaria patients, provided samples and helped with experimental design; P.R. provided computation resources and helped with analysis; K.C. helped with lineage structure algorithm optimization and lineage visualization; N.J. conceived the idea, designed the study, and directed data analysis; B.S.W. and N.J. wrote the paper with contributions from all co-authors.

Here, we use MIDCIRS (See **Chapter 2**) to examine the antibody repertoire diversification in infants (<12 months old) and toddlers (12 – 47 months old) from a malaria endemic region in Mali before and during acute *Plasmodium falciparum* infection. Although the antibody repertoire in fetuses[5], cord blood[6], young adults[7], and the elderly[7, 8] has been studied, infants and toddlers are among the most vulnerable age groups to many pathogenic challenges, yet their immune repertoires are not well understood. Infants are widely thought to have weaker responses than toddlers to vaccines because of their developing immune systems[9]. Thus, understanding how the antibody repertoire develops and diversifies during a natural infection, such as malaria, not only provides valuable insight into B cell ontology in humans, but also provides critical information for vaccine development for these two vulnerable age groups. Using peripheral blood mononuclear cells (PBMC) from 13 children aged 3 to 47 months old before and during acute malaria, with two of the children followed for a second year and 9 additional pre-malaria individuals we show that infants and toddlers use the same V, D, and J combination frequencies and have similar complementarity determining region 3 (CDR3) length distributions. Although infants have a lower level of average SHM than toddlers, the number of SHMs in reads that mutated in infants is unexpectedly high. Infants have a similar, if not higher, degree of antigen selection strength, assessed by the likelihood of amino acid-changing SHMs, compared with toddlers. Remarkably, during acute malaria, antibody lineages expand in both infants and toddlers, and this expansion is coupled with extensive diversification to the same degree as in young adults in response to acute malaria[10, 11]. Furthermore, informatically reconstructing antibody clonal lineages using sequences from both pre-malaria and acute malaria samples from the same individuals shows that infants are capable of introducing SHMs upon a natural infection. This two-timepoint-shared lineage analysis reveals that memory B cells isolated from

pre-malaria samples in malaria-experienced individuals continue to induce SHMs upon acute malaria rechallenge and most IgM memory B cells maintain IgM, while a small fraction switch isotypes. In summary, using an accurate and high-coverage IR-Seq method, we discover features of the antibody repertoire that were previously unknown in infants and toddlers, shedding light on the development of the immune system and its interactions with pathogens.

3.2 RESULTS

3.2.1 Infants and toddlers have similar VDJ usage and CDR3 lengths

Equipped with this ultra-accurate and high-coverage antibody repertoire sequencing tool, we applied it to study the antibody repertoire of infants and toddlers residing in a malaria endemic region of Mali. From an ongoing malaria cohort study[12], we obtained paired PBMC samples collected before and during acute febrile malaria from 13 children aged 3 to 47 months old (**Figure 3.1** and **Table B.1**). Two of the children were followed for an additional year, giving 15 total paired PBMC samples. An average of 3.8 million PBMCs per sample were directly lysed for RNA purification. All PBMCs were subjected to MIDCIRS analysis. An average of 3.75 million sequencing reads were obtained for each PBMC sample (**Table B.2**).

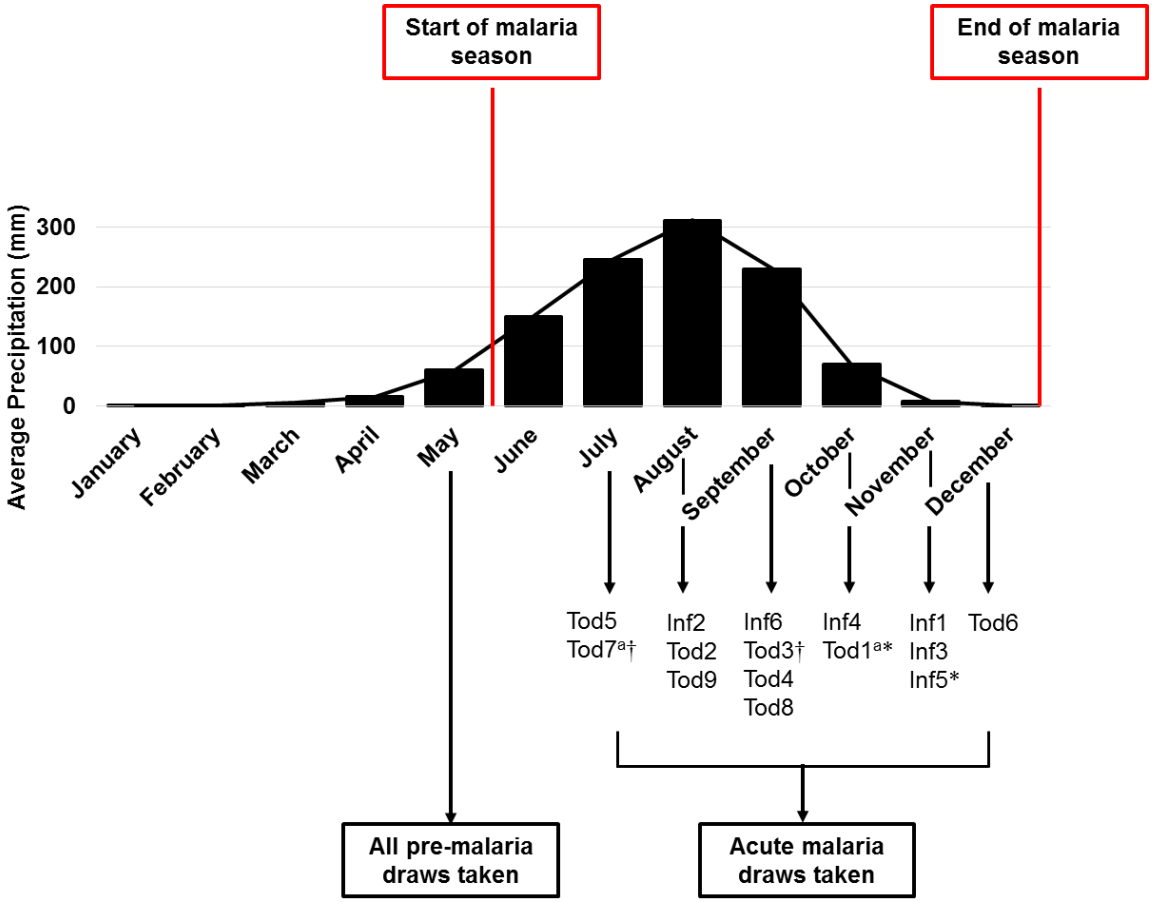

Figure 3.1: Sample collection timeline. All pre-malaria blood draws were taken in May, just before the start of the rainy season. Acute malaria blood draws were taken 7 days after the onset of acute febrile malaria. Unless otherwise indicated ([a]), all samples were collected during 2011. Average precipitation was estimated from the neighboring city of Bamako, Mali (climatemps.com). * Same individual. † Same individual. [a] Drawn in 2012

For all PBMC samples, sequencing approximately the same number of reads as the cell numbers saturates the rarefaction curves (**Figure B.1**). VDJ gene usage is highly correlated for IgM between infants and toddlers regardless of weighting the correlation coefficient by the number of sequencing reads or clonal lineages (**Figure 3.2**), demonstrating that the same mechanism of VDJ recombination is used to generate the primary antibody repertoire in infants and toddlers. Weighting on the number of clonal

lineages in each VDJ class increases the correlation for IgG and IgA compared with weighting on the number of reads in each VDJ class (**Figure 3.2**). The diagonal lines in each panel indicate same sample self-correlation, and the two shorter off-diagonal lines indicate correlations from two timepoints of the same individual. These data recapitulate previous observations from our study in zebrafish that clonal expansion-induced differences on the number of reads in each VDJ class can confound the highly similar VDJ usage during B cell ontology[13]. In addition, infants and toddlers have similar CDR3 length distributions across the three isotypes and both timepoints (**Figure 3.3**), consistent with recent studies of PBMCs from 9 month olds infants[5, 6] and adults[14, 15] and confirming the previous results that an adult-like distribution of CDR3 length is achieved around two months of age[16].

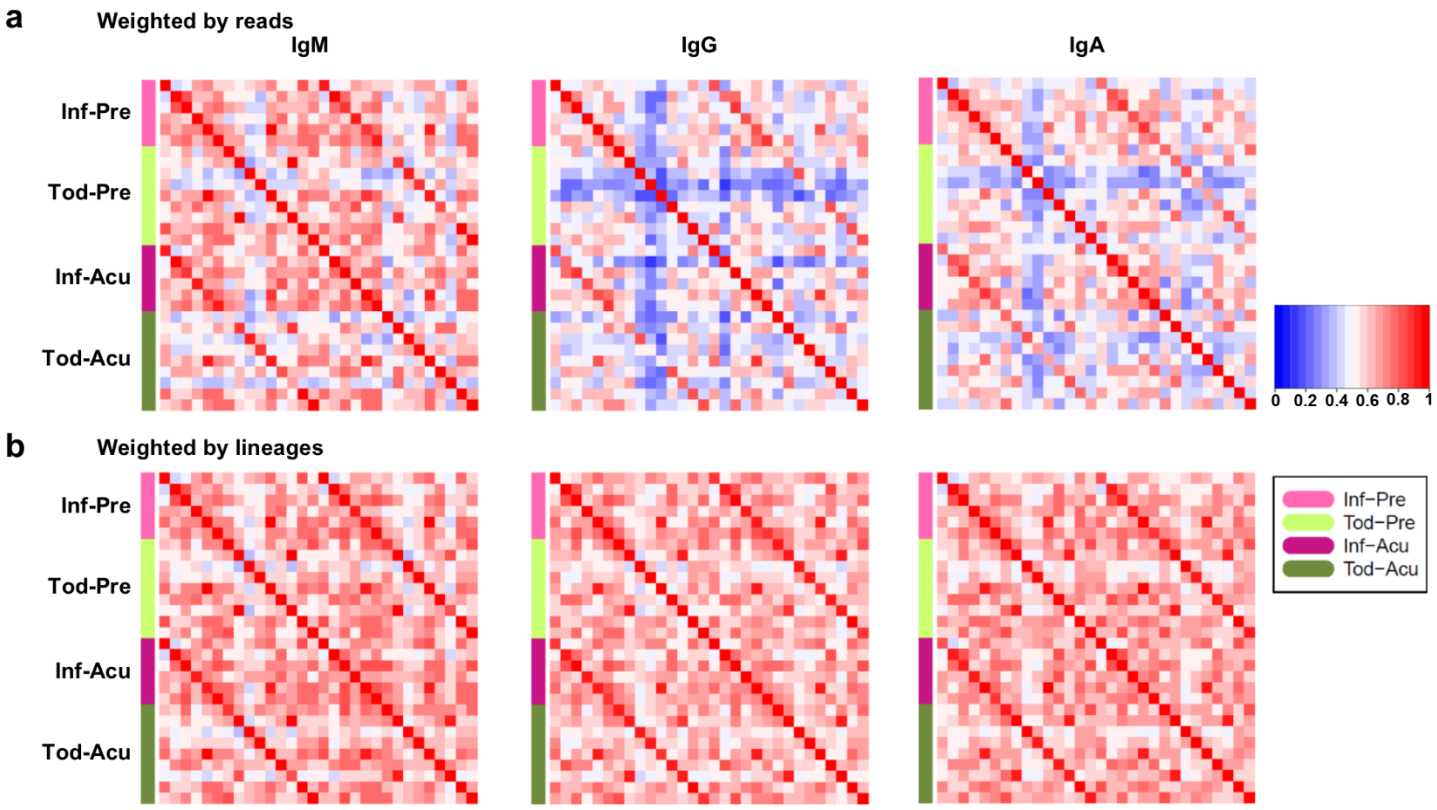

Figure 3.2: Correlation between VDJ usage in paired PBMCs samples (N=15 pairs of pre-malaria and acute malaria). Correlations weighted by reads (**a**) or by lineage (**b**). The color bar left of each panel as well as in figure legend indicates the sample group: infant pre-malaria (pink), toddler pre-malaria (light green), infant acute malaria (maroon), and toddler acute malaria (dark green). Color indicates strength of Pearson correlation. The diagonal lines in each panel indicate same sample self-correlation; two shorter off-diagonal lines indicate correlations from two timepoints of the same individual.

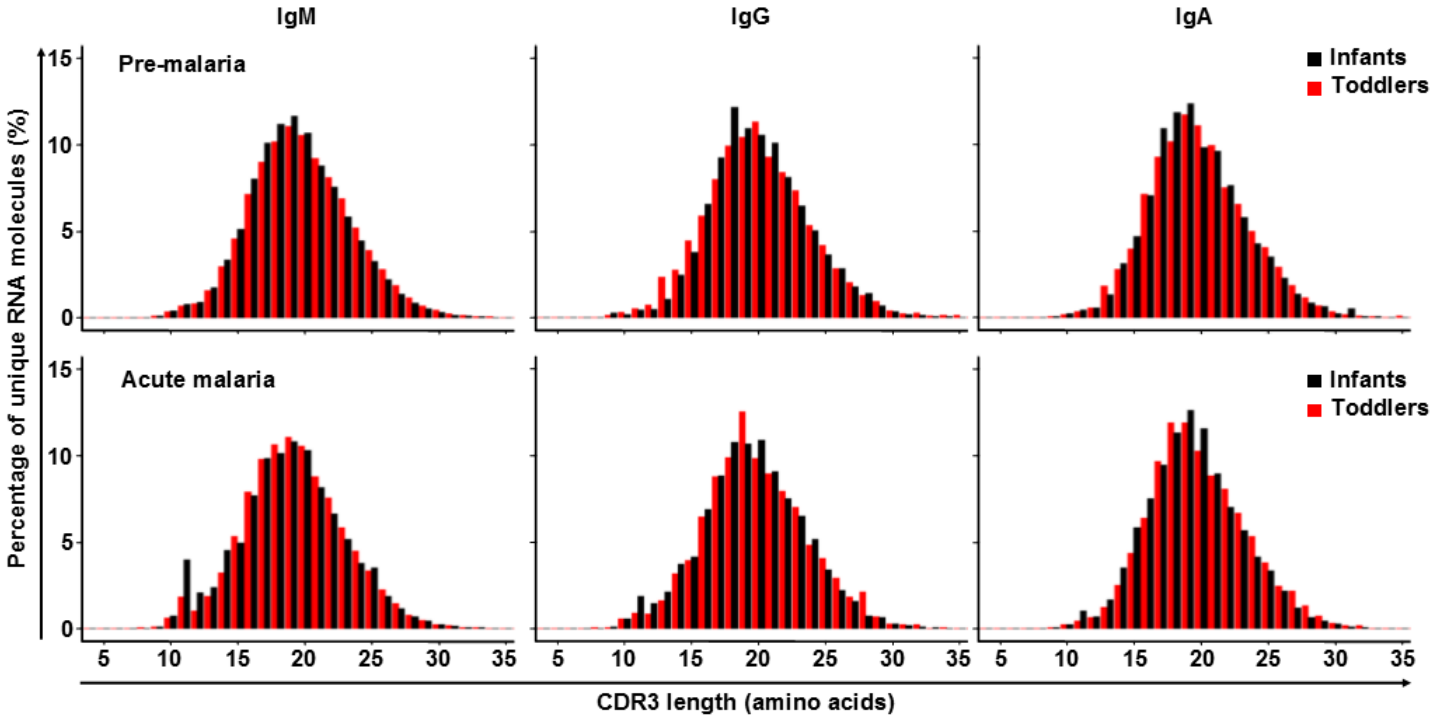

Figure 3.3: Infants and toddlers have similar CDR3 length distributions. CDR3 amino acid lengths of infants (black, N=6) and toddlers (red, N=9) at pre-malaria (top) and acute malaria (bottom) timepoints, separated by isotype.

3.2.2 Both infants and toddlers have unexpectedly high SHM loads

SHM is an important characteristic of antibody repertoire secondary diversification due to antigen stimulation[17]. Although it has been demonstrated before that infants have fewer mutations in their antibody sequences than toddlers and adults, the limited number of sequences for only a few V genes does not provide convincing evidence of the levels of SHM in infants[18]. A recent study using the first generation of IR-Seq showed that two 9-month-old infants averaged at least 6 SHMs in IgM of an average length of 500 nucleotides[5]. These numbers are equivalent to, if not higher than, reported SHM rates in IgM sequences from healthy adults day 7 post influenza vaccination[19] and are much higher than a low-throughput infant study using a few V

genes and limited antibody sequences[20]. Due to inherent errors associated with the first generation of IR-Seq as discussed above, it is possible that PCR and sequencing errors played a role[5]. In addition, it remains unclear if infants (< 12 months old) are able to generate a significant number of mutations in response to infection, which would demonstrate their capacity to diversify the antibody repertoire[21].

Here, we show that infants (< 12 months old) and toddlers (12 – 47 months old) reach an unexpectedly high level of SHMs in all 3 major isotypes, particularly IgG and IgA[22] (**Figure 3.4**). While the mutation distributions remain in the low end of the spectrum for IgM, the number of mutations is significantly higher in IgG and IgA for both age groups. The threshold for the 10% most highly mutated unique RNA molecules is around 10 in infant IgG and IgA sequences (**Figure 3.4**, Infants, right of the blue long vertical lines) and around 20 in toddler IgG and IgA sequences (**Figure 3.4**, Toddlers, right of the blue long vertical lines). To minimize any possible inflation of SHMs, we excluded all sequences that were mapped to novel alleles, which were identified by both TIgGER[23] and inspecting IgM sequences (see **Chapter 4**). These putative novel alleles account for 8% of all unique sequences on average (**Table B.3**). Naive B cells from these same patients, sorted as a control, harbor only 0.55 mutations on average, as expected (**Table B.4**). Upon acute malaria infection, the SHM histogram shifts rightward for almost all isotypes in almost all individuals (**Figure 3.4**, the right shift of pink long vertical line compared to blue long vertical line), including infants. These results demonstrate high levels of SHM that exceed what have been documented previously[20, 22].

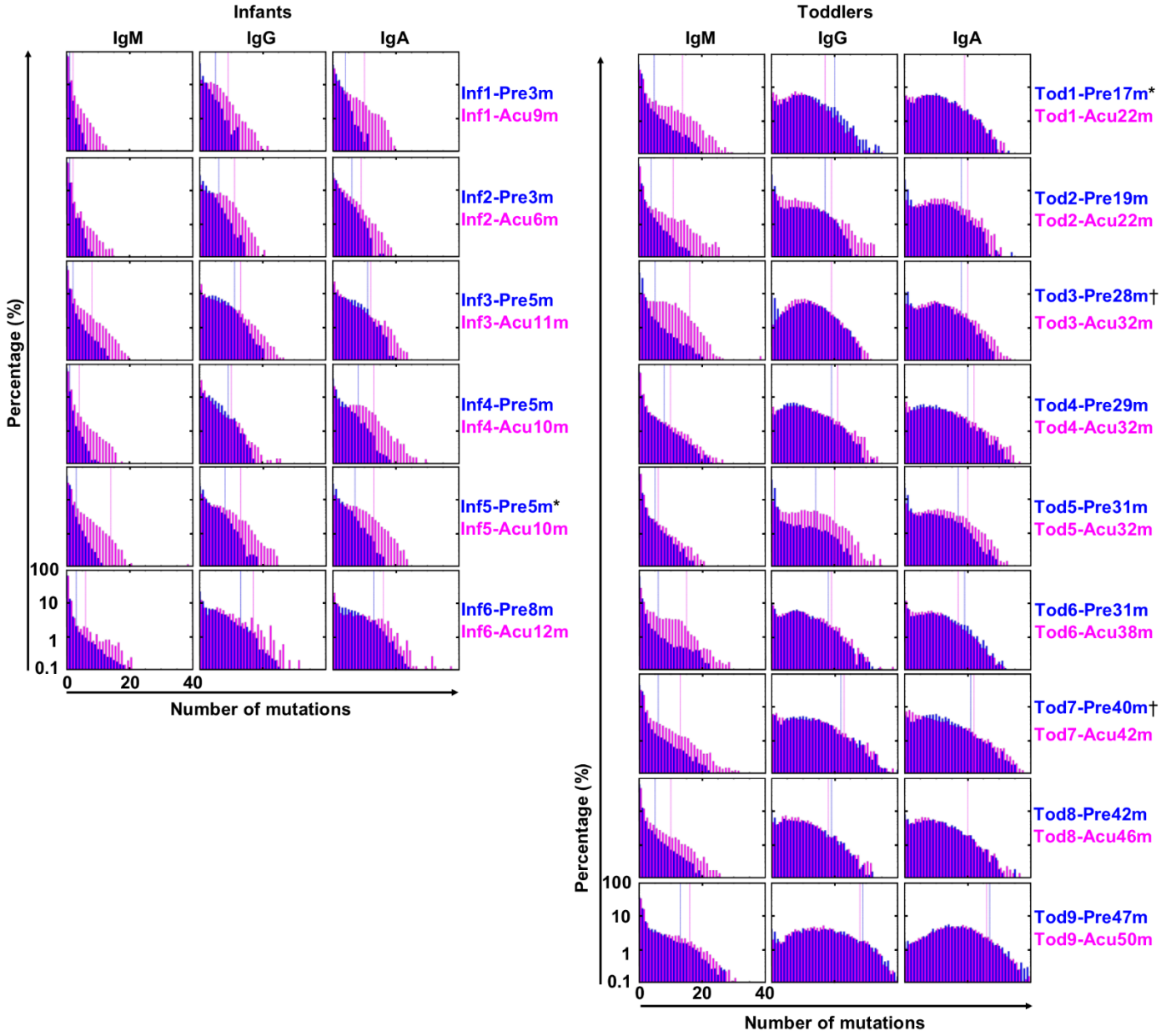

Figure 3.4: Infants and toddlers are capable of generating highly mutated antibodies. Distribution of SHM number for infants (N=6) and toddlers (N=9), from whom we had paired pre-malaria (blue) and acute (pink) malaria samples, weighted by unique RNA molecules. Blue and pink long vertical lines represent the number of mutations above which 10% of sequences fall for the respective samples. * and † demarcate samples derived from the same individuals followed for 2 malaria seasons.

3.2.3 SHM load is distinct between infants and toddlers

The differences in the shapes of SHM distributions of infants and toddlers, steadily decreasing from unmutated for infants in all three isotypes while peaking around 10 for toddlers in IgG and IgA (**Figure 3.4**), suggest that the total SHM load might reflect the history of interactions between the antibody repertoire and the environment, including malaria exposure. Since the malaria season is synchronized with the 6-month rainy season (**Figure 3.1)**, and > 90% of the individuals in this cohort are infected with *P. falciparum* during the annual malaria season[12], we hypothesized that the SHM load would increase with age. However, we found that the SHM load rapidly increases with age in infancy and then appears to plateau around 12 months of age in an initial smaller set of children with paired pre-malaria and acute malaria PBMC samples (**Figure B.2**). We then added 9 pre-malaria samples around the infant and toddler transition (5 of 11 months old and 4 of 13 to 17 months old). The two-staged trend of SHM load remains for all three isotypes (**Figure 3.5**), with samples around the transition having the largest variation. Detailed comparisons show that, consistent with the two-stage trend, toddlers have a higher SHM load compared with infants for all three isotypes at both pre-malaria and acute malaria timepoints (**Figure 3.6**, comparison between age groups). Although there is a significant increase on SHM load upon acute malaria infection in IgM for both infants and toddler, bulk PBMC analysis does not show a significant increase in IgG or IgA, possibly because of the already elevated SHM base level. This, along with the two-stage trend (**Figure 3.5**), suggests that 12 months is an important developmental threshold for secondary antibody repertoire diversification: before this threshold, the global repertoire is quite naive but can quickly diversify upon a natural infection.

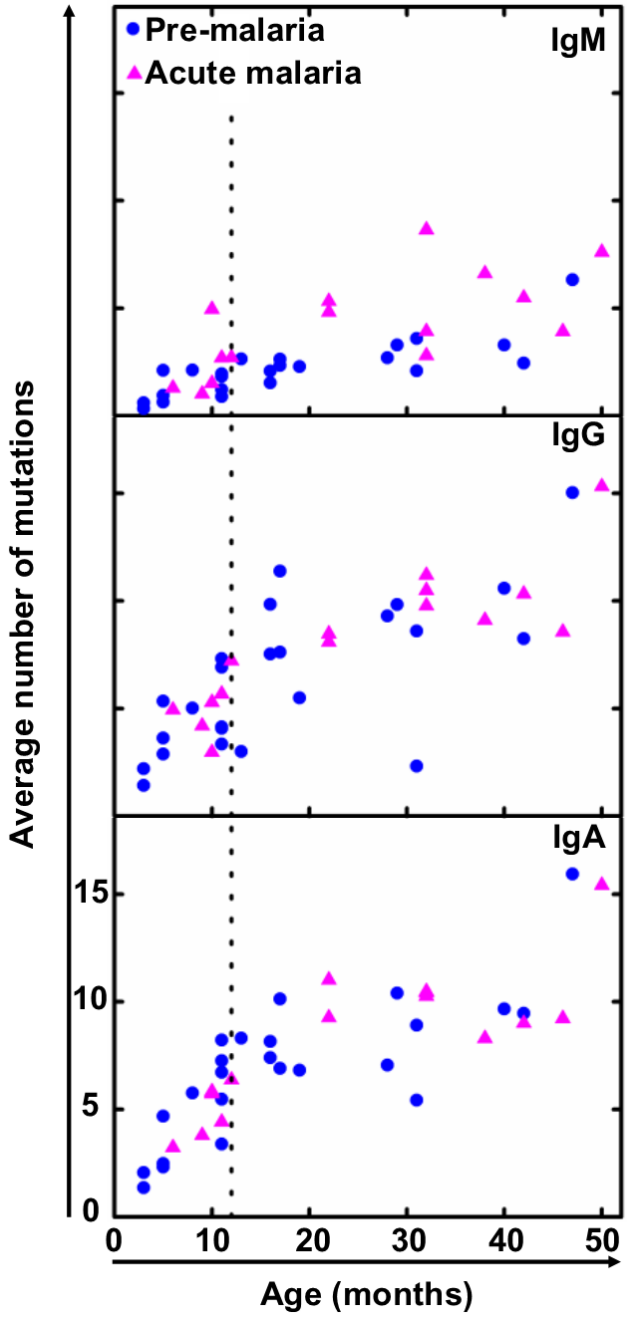

Figure 3.5: SHM load increases rapidly with age before reaching a plateau. Age-related average number of mutations in pre- (blue circle, N=24, N_{Infant}=11, $N_{Toddler}$=13) and acute malaria (pink triangle, N=15, N_{Infant}=6, $N_{Toddler}$=9) samples, weighted by RNA molecules, split by isotype. Dashed line indicates the age boundary for infants (<12 months old) and toddlers (12 – 47 months old).

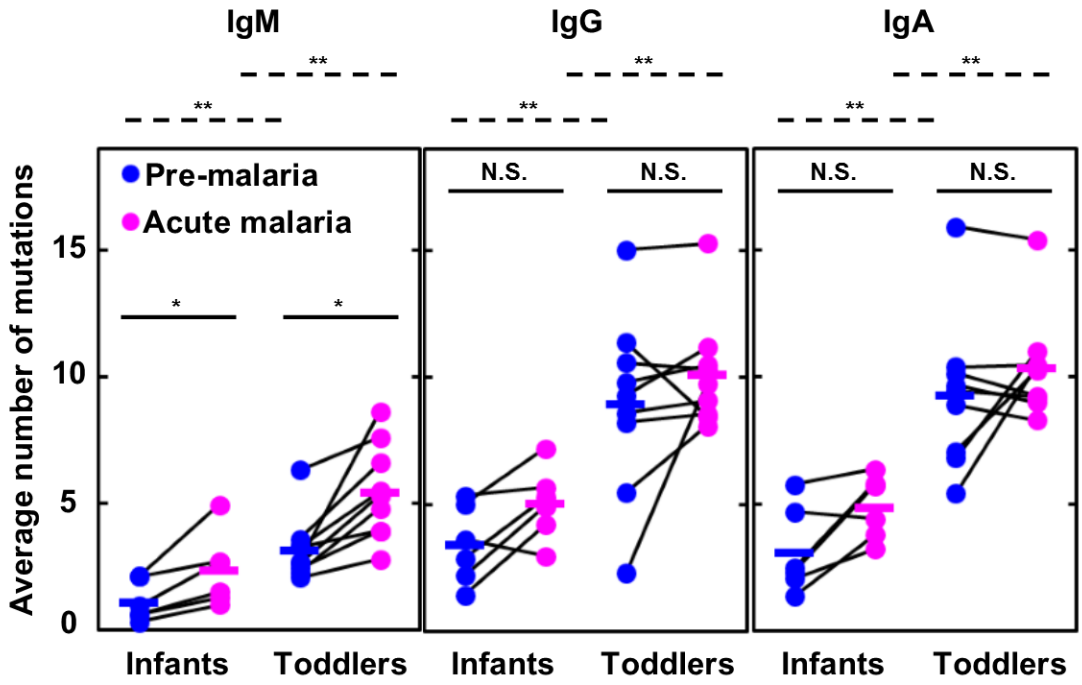

Figure 3.6: Comparison of average number of mutations for paired infants and toddlers. Pre- (blue) and acute (pink) malaria samples separated by isotype; lines connect paired samples ($N_{Infant,paired}$=6, $N_{Toddler,paired}$=9). Bars indicate means. *$P < 0.05$, **$P < 0.01$, N.S. indicates no significant difference by two-tailed Mann-Whitney U test (between age groups, dashed lines) or two-tailed Wilcoxon Signed-Rank test (between paired timepoints, solid lines). Differences in variance were not significant by squared ranks test.

3.2.4 Higher memory B cell percentage results in higher SHM load

This unexpected developmental threshold of secondary antibody repertoire diversification prompted us to focus on B cell subset composition changes and ask whether they correlate with this two-staged SHM load. Flow cytometry analysis reveals that naive B cells decrease from about 95% in 3-month-old infants to about 80% in toddlers (**Figure 3.7a**). Conversely, memory B cells increase from about 4% in 3-month-old infants to about 15% in toddlers (**Figure 3.7f**). As the two-stage SHM load analysis suggests, 12 months appears to divide the samples into two age groups, with a large

variation at the infant to toddler transition and in the toddler group. Infants have a significantly more naive B cells and fewer memory B cells than toddlers (**Figure 3.7b,g**). Plasmablast percentages fluctuated in a much smaller range (**Figure B.3**). With a similar two-staged trend observed for B cell subset percentages, we hypothesized that the B cell subset percentage would correlate with SHM load. Indeed, further analysis shows that the decrease in naive B cell percentage and the increase in memory B cell percentage correlate well with SHM load across IgM, IgG, and IgA isotypes (**Figure 3.7c-e** and **h-j**), which supports our initial hypothesis that 12 months separates infants from toddlers in both SHM load and B cell composition changes. These data suggest that memory B cells contribute significantly to the developing antibody repertoire, and their composition is essential in secondary antibody repertoire diversification.

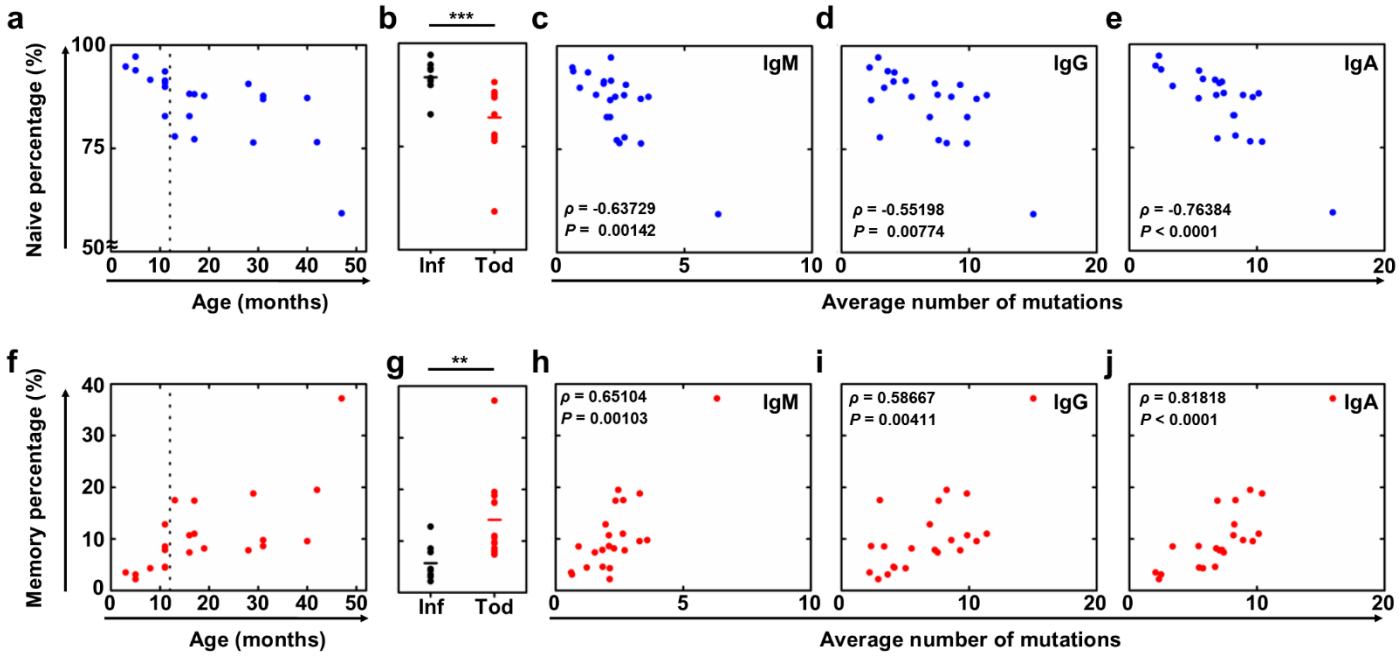

Figure 3.7: Decrease of naive B cell and increase of memory B cell percentages show a two-stage trend and correlate with SHM load. (**a**) Naive B cell percentages of total B cells from the pre-malaria samples (N=22) vary with age. Dashed vertical line depicts the cutoff between infants and toddlers. (**b**) Naive B cell percentages of total B cells compared between infants (black, N=9) and toddlers (red, N=13). (**c-e**) Naive B cell percentages correlate with average number of mutations (SHM load) in IgM (**c**), IgG (**d**), and IgA (**e**) sequences from bulk PBMCs in pre-malaria samples (N=22). (**f**) Memory B cell percentages of total B cells from the pre-malaria samples (N=22) vary with age. Dashed vertical line depicts the cutoff between infants and toddlers. (**g**) Memory B cell percentages of total B cells compared between infants (black, N=9) and toddlers (red, N=13). (**h-j**) Memory B cell percentages correlate with average number of mutations (SHM load) in IgM (**h**), IgG (**i**), and IgA (**j**) sequences from bulk PBMCs in pre-malaria samples (N=22). (**b** and **g**) Bars indicate means; $^{**}P < 0.01$, $^{***}P < 0.001$, two-tailed Mann-Whitney U test. (**c** to **e** and **h-j**) ρ and P values determined by Spearman's rank correlation listed in each panel.

3.2.5 SHMs are similarly selected in infants and toddlers

One of the key features of antibody affinity maturation is antigen selection pressure imposed on an antibody, which is reflected in the enrichment of replacement mutations[24] in the CDRs, the parts of the antibody that interact with antigens, and the depletion of replacement mutations in the framework regions (FWRs), the parts of the antibody responsible for proper folding. The unexpectedly high level of SHMs observed in infants prompted us to ask whether those SHMs have characteristics of antigen selection, as seen in older children and adults. As previous studies have shown that infants have limited CD4 T cell responses and neonatal mice exhibit poor germinal center formation[9], we hypothesized that infant antibody sequences would display weaker signs of antigen selection. Here, we use a recently published tool, BASELINe[25], to compare the selection strength. BASELINe quantifies the likelihood that the observed frequency of replacement mutations differs from the expected frequency under no selection; a higher frequency implies positive selection and a lower frequency implies negative selection, and the degree of divergence from no selection relates to the selection strength. Surprisingly, despite infants harboring fewer overall mutations, these mutations are positively selected in the CDRs and negatively selected in the FWRs in both IgG and IgA (**Figure 3.8b,c,e,f**). Contrary to the hypothesis that infants would have a lower selection strength than toddlers, for both IgG and IgA, infants actually have a higher selection strength at both pre-malaria and acute malaria timepoints (**Figure 3.8**). The lower selection strength in infant IgM sequences at the pre-malaria timepoint is significantly higher during acute malaria infection (**Figure 3.8a,d,** CDR black curves between two timepoints, $P < 0.0001$ [numerical integration, as previously described[25]]), suggesting that the significant increase in SHM is antigen-driven and selected upon. In order to compare with a large amount of historical adult data, we calculated replacement to silent

mutation ratios (R/S ratios), which are about 2-3:1 in FWRs and 5:1 in CDRs for both infants and toddlers (**Table B.5**). These results are similar to adults[24, 26, 27, 28] and much higher than what has been reported for children previously using a very limited number of sequences[29]. We also noticed that R/S ratio in the FWRs of IgM was much higher in infants, contrary to the BASELINe results, which highlights the importance of incorporating the expected replacement frequency when considering selection pressure. These results suggest that as an end result of interactions between antigen selection and SHM, the degree of antibody amino acid changes is comparable in infants, toddlers, and adults. It also suggests that cellular and molecular machineries for antigen selection are already in place in infants.

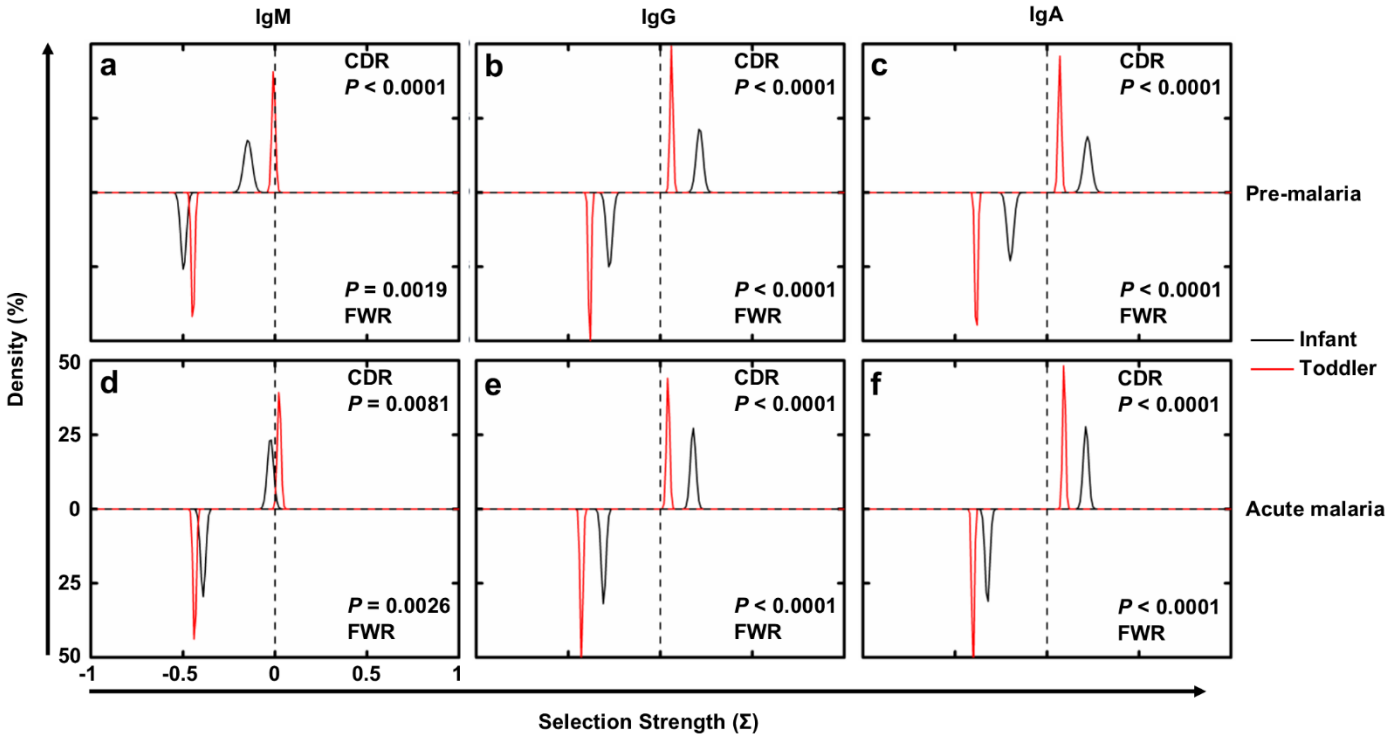

Figure 3.8: Antigen selection strength comparisons between infants and toddlers. Selection strength distributions, as determined by BASELINe[25], were compared between infants (black) and toddlers (red) for PBMCs from pre-malaria (**a-c**) (N_{infant}=6, $N_{toddler}$=9) and acute malaria (**d-f**) (N_{infant}=6, $N_{toddler}$=9) timepoints, separated by isotype: (**a,d**) IgM, (**b,e**) IgG, and (**c,f**) IgA. Selection strength on CDR (CDR1 and 2, top half of each panel) and FWR (FWR2 and 3, bottom half of each panel) for unique RNA molecules was calculated. CDR3 and FWR4 were omitted due to the difficulty in determining the germline sequence. FWR1 for all sequences was also omitted because it was not covered entirely by some of the primers. *P* value calculated as previously described[25].

3.2.6 Clonal lineages diversify upon acute febrile malaria

The exhaustive sequencing data obtained by MIDCIRS offers the possibility to reconstruct clonal lineages that trace B cell development. Clonal lineages contain different species of unique antibody sequences that could be progenies derived from the same ancestral B cell. B cell clonal lineage analysis has been used to track affinity

maturation and sequence evolution of HIV broadly neutralizing antibodies[30, 31]. Using a clustering method with a pre-determined threshold (90% similarity on nucleotide sequence at CDR3), we previously demonstrated that B cell clonal lineages could be informatically defined and contain pathogen-specific antibody sequences[7]. In addition, the clonal lineage analysis also highlighted the lack of antibody diversification in the elderly after influenza vaccination[7]. Using the same approach and a similar threshold[7, 32], we aimed to answer whether infants and toddlers are able to diversify antibody clonal lineages in response to infection and, if so, whether they have a similar ability to do so, which was previously impossible to answer due to technical limitations. To do this, we first visualized structures of informatically defined clonal lineages for the entire antibody repertoire (**Figure B.4**). Each oval lineage map represents an individual PBMC sample at one timepoint. Densely packed individual lineages are not easily identified visually in **Figure B.4**; however, dark areas indicate that clonal lineages are already complex in this cohort of infants as young as 3 months old and can be further diversified upon acute febrile malaria.

The densely packed lineages could result from large lineage sizes (one unique RNA molecule with many copies), large lineage diversities (many unique RNA molecules), or a combination of the two. To closely examine the possible differences in the degree of this intra-clonal lineage expansion and diversification between infants and toddlers, especially upon acute febrile malaria, we projected the global lineage structure (**Figure B.4**) onto diversity and size of lineage axes (**Figure 3.9**). Each circle represents an individual lineage, with the area of the circle proportional to the SHM load (average mutations of the lineage). This analysis effectively captures five parameters that quantify lineage complexity in a sample: number of total clonal lineages (number of circles), diversity of each lineage (x-axis position, number of unique RNA molecules in a

lineage), size of each lineage (y-axis position, number of total RNA molecules in a lineage), SHM load of each lineage (area of circle, key is located in between the infant and toddler panels in **Figure 3.9**), and the extent of clonal expansion of each lineage (distance from y=x parity line; no clonally expanded RNA molecules within a lineage if it is on parity line or pure clonal expanded RNA molecules if it is in the top left quadrant of each panel).

Figure 3.10 shows two example lineages selected to display the full lineage structures to demonstrate a lineage with diversification and clonal expansion (**Figure 3.10a** refers to green letter "a" indicated in **Figure 3.9**, Inf3) and another one with diversification but without clonal expansion (**Figure 3.10b** refers to green letter "b" indicated in **Figure 3.9**, Inf3). Both are represented by a single circle in **Figure 3.9**, but their locations in **Figure 3.9** depend on the numbers of RNA molecules (y-axis) and numbers of unique RNA molecules (x-axis). Lineage "b" (b in **Figure 3.9**, Inf3, zoomed in view in **Figure 3.10b**) that lies away from the origin and near the black y=x parity line consists of 8 unique sequences, each represented by only one RNA molecule, indicating extensive lineage diversification but no clonal expansion. Lineage "a" (a in **Figure 3.9**, Inf3, zoomed in view in **Figure 3.10a**) that lies far from the parity line is dominated by two unique RNA molecules each with about 20 copies (**Figure 3.10a**, height of nodes), indicating extensive clonal expansion of particular sequences in addition to diversification. Changing lineage forming threshold from 90% to 95% does not change the overall structure of the lineages (**Figure B.5**).

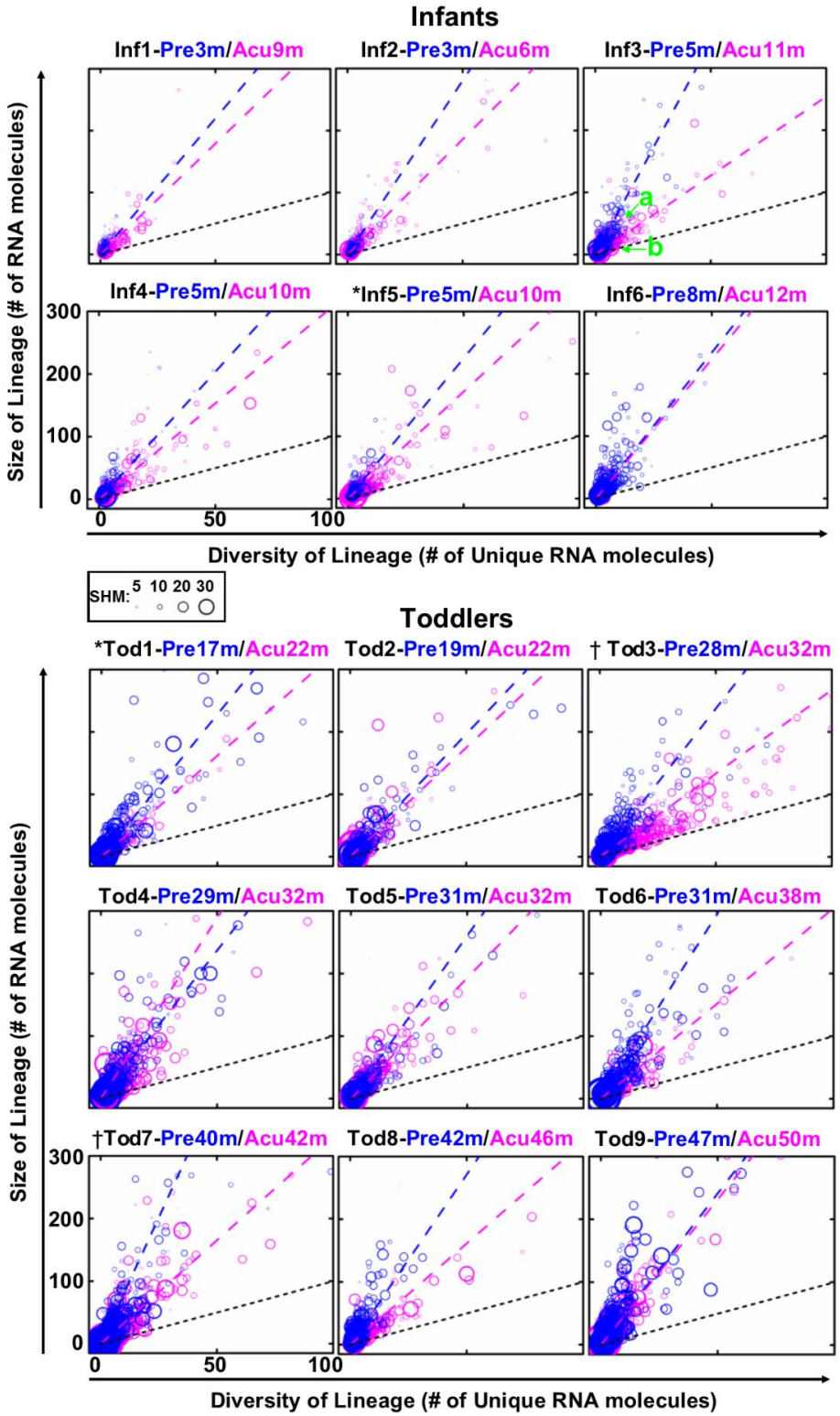
Infants
Inf1-Pre3m/Acu9m
Inf2-Pre3m/Acu6m
Inf3-Pre5m/Acu11m
a
b
Inf4-Pre5m/Acu10m
*Inf5-Pre5m/Acu10m
Inf6-Pre8m/Acu12m
Size of Lineage (# of RNA molecules)
300
200
100
0
0
50
100
Diversity of Lineage (# of Unique RNA molecules)
SHM: 5 10 20 30
Toddlers
*Tod1-Pre17m/Acu22m
Tod2-Pre19m/Acu22m
† Tod3-Pre28m/Acu32m
Tod4-Pre29m/Acu32m
Tod5-Pre31m/Acu32m
Tod6-Pre31m/Acu38m
†Tod7-Pre40m/Acu42m
Tod8-Pre42m/Acu46m
Tod9-Pre47m/Acu50m
Size of Lineage (# of RNA molecules)
300
200
100
0
0
50
100
Diversity of Lineage (# of Unique RNA molecules)

Figure 3.9

Figure 3.9: B cell lineage complexity change under malaria stimulation. Diversity and size of B cell lineages for infants (N=6) and toddlers (N=9) from whom paired PBMC samples at pre-malaria (blue) and acute malaria (pink) were obtained. Each circle represents an individual lineage. The area of each circle is proportional to the SHM load. Labeled green arrows indicate representative lineages whose intra-lineage structures were shown in detail in **Figure 3.10**. Each circle's x and y coordinates were determined by its diversity (the number of unique RNA molecules in a lineage) and size (the number of total RNA molecules in a lineage), respectively. Blue and pink dashed lines represent the linear fit for pre- and acute malaria lineages, respectively. Black dashed lines indicate y=x parity, such that lineages lying on the parity line are comprised entirely of unique RNA molecules with minimum clonal expansion, such as lineage in **Figure 3.10b**. On the other hand, lineages comprised of clonally expanded RNA molecules are close to the y axis, such as lineage **Figure 3.10a**.

This five-dimension lineage analysis reveals that infants as young as 3 months old can generate extensive lineage structures, with many lineages containing more than 20 different types of antibody sequences and 50 RNA molecules (**Figure 3.9**). Toddlers have many more lineages with higher levels of both size and diversity. However, in both infants and toddlers, the majority of clonal lineages are singleton lineages consisting of only one RNA molecule (**Figure 3.11a**), consistent with the flow cytometry analysis that the bulk of the B cell repertoire is naive in these young children (**Figure 3.7**). Upon acute malaria infection, the fraction of non-singleton lineages increases in both infants and toddlers (**Figure 3.11a**).

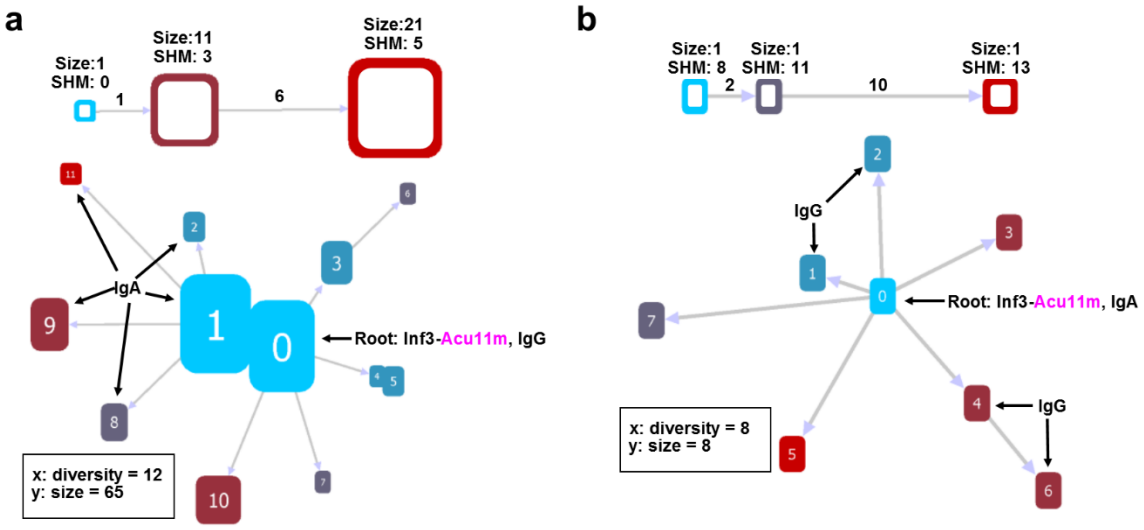

Figure 3.10: Visualized example lineages. (**a**) corresponds to green "a" labeled lineage in **Figure 3.9**. (**b**) corresponds to green "b" labeled lineage in **Figure 3.9**. Each node is a unique RNA molecule species. The height of the node corresponds to the number of RNA molecules of the same species, the color corresponds to number of nucleotide mutations, and the distance between nodes is proportional to the Levenshtein distance between the node sequences, as indicated in the legend above each lineage. All unlabeled nodes share the isotype with the root.

In order to tease out whether these non-singleton lineages diversify or clonally expand upon acute infection, we fit linear regressions to the lineage diversity-size plots. An immune response against an infection can have a two-fold effect on the lineage landscape: antigen stimulation can cause clonal expansion, which would shift the lineage up on the y-axis, and SHM and affinity maturation, which would shift the lineage to the right on the x-axis. This balance between clonal expansion and diversification is depicted by the slope of the linear regression (**Figure 3.9**, dashed blue lines for pre-malaria samples and dashed pink lines for acute malaria samples). We hypothesized that the lower absolute SHM load of infants would imply a defect in the ability to diversify clonal lineages in response to infection, leading the slope change from pre-malaria to acute

malaria to be low (a small angle between blue and pink dashed lines) or even negative (pink dashed line is closer to y-axis than blue dashed line). Surprisingly, our analysis shows that infants diversify their clonal lineages in a similar manner as toddlers in response to acute malaria (**Figure 3.11b**). As singleton lineages do not bear any weight on the linear regression, our analysis shows that the increasing fraction of non-singleton lineages upon malaria infection is similarly diversified between infants and toddlers, which is also similar to a young adult at pre-malaria and acute malaria (**Figure B.6**). However, this sharply contrasts with what we had previously observed in the elderly following influenza vaccination, where clonal expansion dominated[7]. Among clonally expanding and diversifying B cell clones during an infection, only a subset of the cells comprising the clonal burst remain once the infection has been cleared. Thus, the characteristic change in the lineage size/diversity linear regression slope upon infection is expected to subside as time passes since the acute infection. Indeed, comparing the pre-malaria lineage size/diversity linear regression slopes reveals no difference between infants (who have not experienced malaria before) and toddlers (who have experienced malarias in previous years) (**Figure B.7**). These results highlight the unexpected capability of young children's antibody repertoire in response to a natural infection.

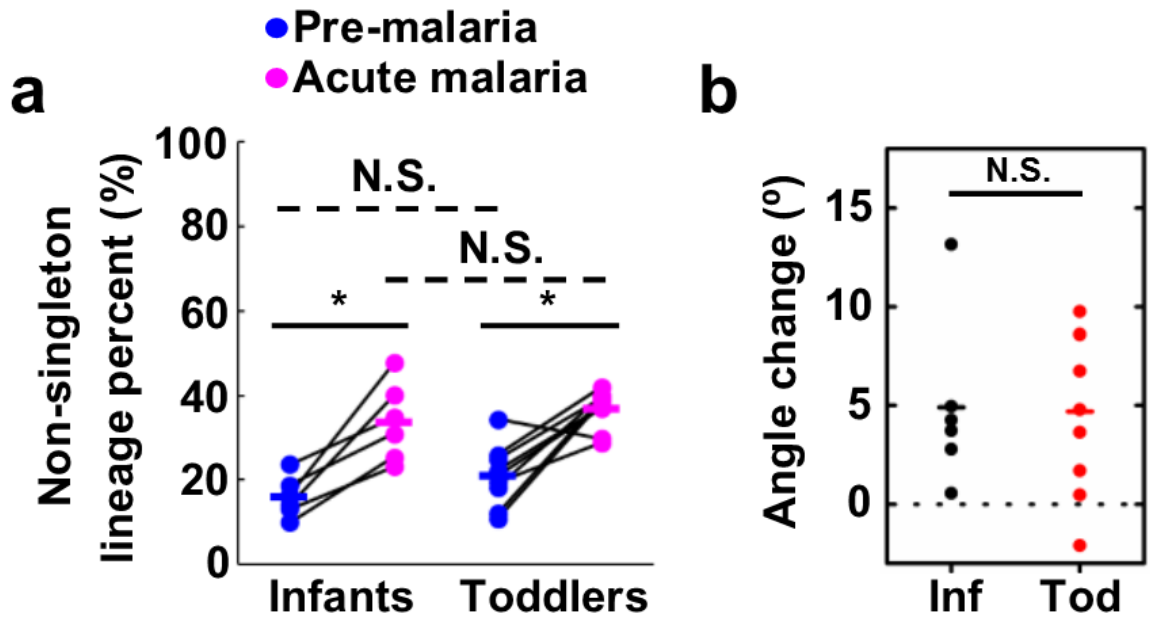

Figure 3.11: Infants and toddlers similarly diversify clonal lineages during acute malaria infection. (**a**) The non-singleton lineage percent (lineages comprised of at least 2 RNA molecules) between infants and toddlers at pre- (blue) and acute (pink) malaria. $*P < 0.05$ by two-tailed Wilcoxon Signed-Rank test (between timepoints, solid lines); N.S. indicates no significant difference by two-tailed Mann-Whitney U test (between age groups, dashed lines). (**b**) The difference of linear regression slopes (angles) from **Figure 3.9**, or degree of diversity change, between pre- and acute malaria for infants (black) and toddlers (red). N.S. indicates no significant difference by two-tailed Mann-Whitney U test. Bars indicate means. Differences in variance were not significant by squared ranks test.

3.2.7 SHM load increases upon acute febrile malaria

The plateau we observed on SHM load in toddlers at both pre- and acute malaria (**Figure 3.5**) and the lack of a SHM difference in IgG and IgA between pre- and acute malaria (**Figure 3.6**) seems to suggest that the experienced part of the repertoire does not respond to malaria infection by inducing SHM. However, it could be that only a portion of the bulk antibody repertoire responds to the infection and there is already a high level of baseline SHMs as revealed by the histogram analysis (**Figure 3.4**). Since we saw the

lineage diversification upon malaria infection in **Figure 3.9**, we hypothesized that examining the SHMs from sequences in two-timepoint-shared lineages (lineages containing both pre-malaria and acute malaria sequences) would enable us to quantify the infection-induced SHM increase from the highly mutated background. To test this, we pooled all sequences from both timepoints, including sorted memory B cells from the pre-malaria timepoint, and generated lineages again using the 90% similarity threshold at CDR3[7, 32]. We are able to find two-timepoint-shared lineages in all individuals analyzed (**Table B.6**). Consistent with the observation that toddlers already have a diverse and expanded antibody repertoire compared to infants, there are more shared lineages in toddlers than infants (**Table B.6**). We tallied SHMs for sequences from pre-malaria and acute malaria in the two-timepoint-shared lineages separately. Consistent with our hypothesis, both infants and toddlers significantly increase SHM upon infection (**Figure 3.12a**). Indeed, toddlers had a higher pre-malaria SHM level compared to infants (**Figure 3.12a**, blue symbols). To our surprise, infants were able to induce more SHMs compared to toddlers (**Figure 3.12b**). These data suggested that indeed both infants and toddlers induce SHMs upon malaria infection.

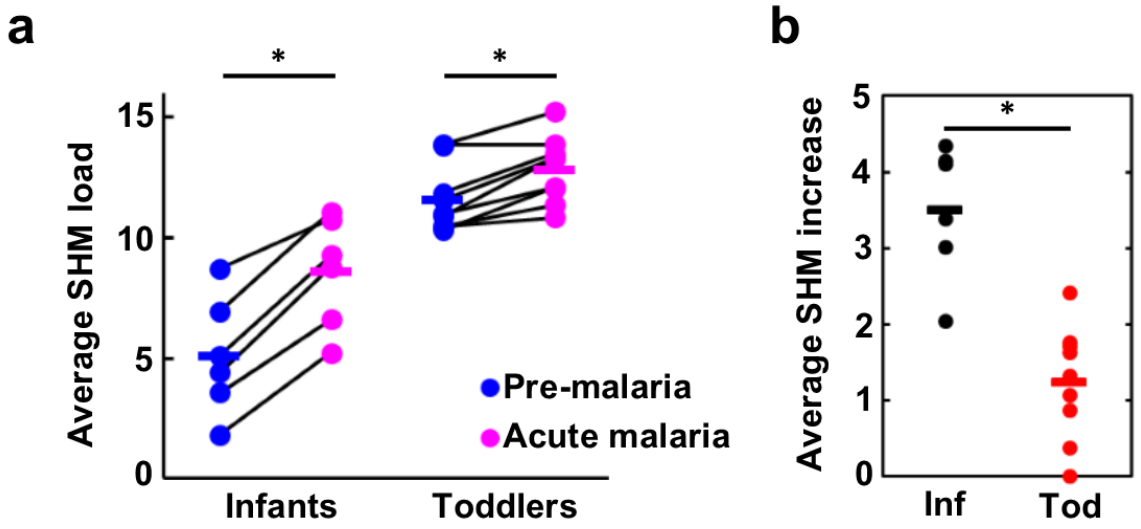

Figure 3.12: Two-timepoint-shared lineage analysis reveals SHM increment during acute malaria infection. (**a**) Average SHM for sequences from pre-malaria (blue) and acute malaria (pink) timepoints within lineages containing sequences from both timepoints for infants (N=6) and toddlers (N=9).) *$P < 0.05$ by two-tailed Wilcoxon Signed-Rank test. (**b**) Average SHM increase upon acute malaria infection for infants (black) and toddlers (red) from (**a**). *$P < 0.05$ by two-tailed Mann-Whitney U test.

3.2.8 Memory B cells further diversify upon malaria rechallenge

The importance of IgM-expressing memory B cells has been reported in mice in several recent studies[33, 34, 35, 36], including a mouse model of malaria infection[37]. However, fewer studies have examined these cells in humans[21], and their composition and role in repertoire diversification upon rechallenge remains elusive. It is widely believed that they may retain the capacity to introduce further mutations and class switch[33, 34, 37, 38]. However, sequence-based clonal lineage evidence is lacking. The paired samples before and during acute malaria from toddlers who experienced malaria in previous years provided an opportunity to investigate the role of memory B cells in repertoire diversification upon rechallenge in children.

Here, we focus on two-timepoint-shared lineages that harbor sequences from pre-malaria memory B cells. Given the significant increase of SHM we identified at acute malaria sequences over pre-malaria sequences in two-timepoint-shared lineages (**Figure 3.12a**), we reasoned that the high repertoire coverage of MIDCIRS should enable us to identify a large number of two-timepoint-shared lineages that contain these memory B cells, and these memory B cells should have mutated progenies at the acute malaria timepoint. To ensure that we identify sequence progenies of these pre-malaria memory B cells, we employed an antibody lineage structure construction algorithm, COLT, that we recently developed[39]. COLT considers isotype, sampling time, and SHM pattern when constructing an antibody lineage, which allows us to trace, at the sequence level, the acute progeny of these memory B cells. As illustrated by **Figure 3.13**, this COLT-generated lineage tree depicts a pre-malaria memory B cell sequence serving as a parent node to sequences derived from the acute malaria timepoint. This analysis is much more stringent in identifying sequence progenies than simply judging if a pre-malaria memory B cell sequence is grouped with acute malaria PBMC sequences.

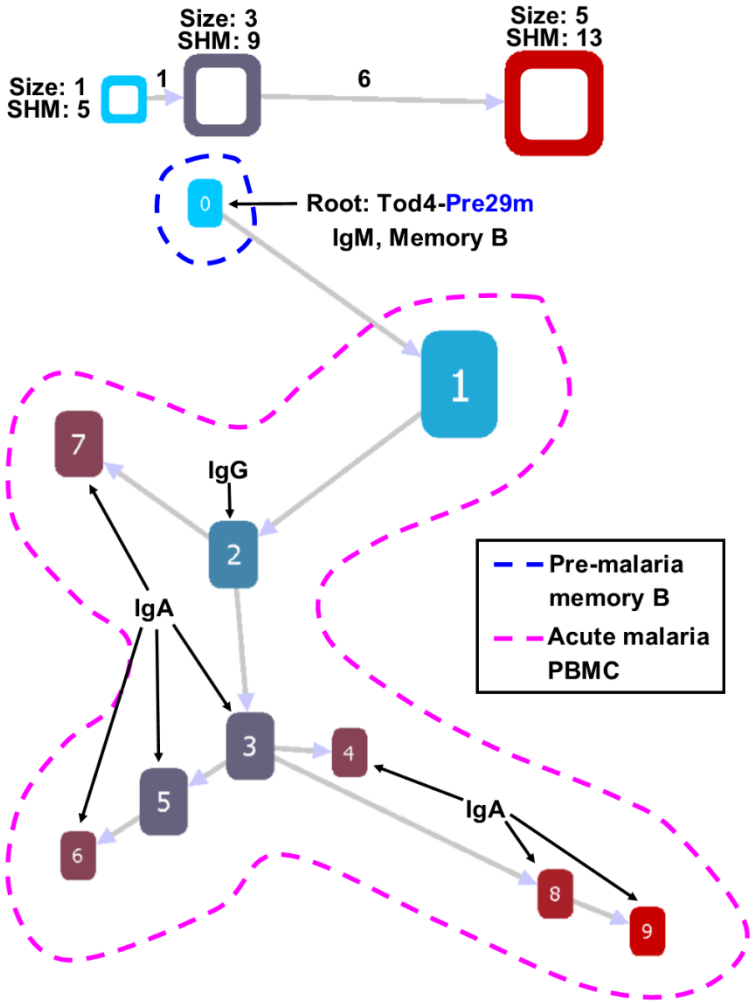

Figure 3.13: Multi-timepoint shared lineage example. Intra-lineage structure for a representative lineage from **Figure 3.14**. Blue dashed curve encompasses the pre-malaria timepoint derived sequence, and pink dashed curve encompasses the acute malaria timepoint derived sequences. Each node is a unique RNA molecule species. The height of the node corresponds to the number of RNA molecules of the same species, the color corresponds to the SHM load, and the distance between nodes is proportional to the Levenshtein distance between the node sequences, as indicated in the legend above the lineage. Unlabeled node shares the isotype with the root.

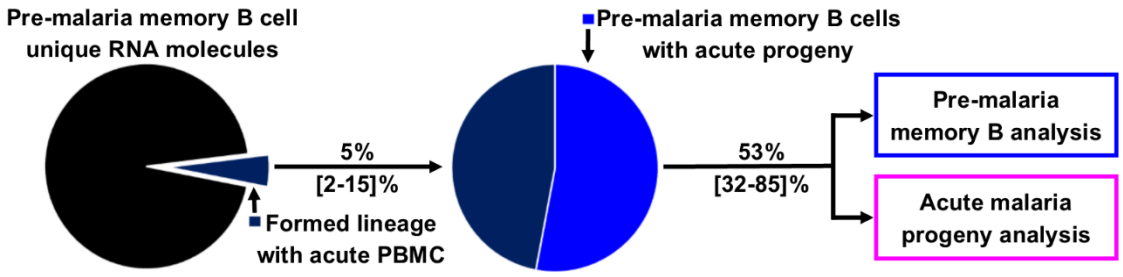

Figure 3.14: Flow diagram for two-timepoint-shared lineage containing pre-malaria memory B cell identification and acute progeny analysis. Percentages represent the average percent of unique sequences classified by the indicated slice, range in brackets.

On average, 5% of unique sequences from 10,000 sorted memory B cells form lineages with acute malaria PBMC sequences (**Figure 3.14**, dark blue slice of the first pie). COLT[39] analysis on these pre-malaria memory B cell-containing lineages shows that 53% contain traceable progeny sequences from the acute malaria PBMCs (**Figure 3.14**, lighter blue slice of the second pie). Overall, there is a significant increase of SHM in these acute malaria progenies compared with their ancestor pre-malaria memory B cells (**Figure 3.15a**). Consistent with previous studies[33, 34, 38], these progeny-bearing pre-malaria memory B cells express all three major isotypes, with IgM being the dominant species (**Figure 3.15b**). Investigating their isotype switching capacity reveals that about 60% of the IgM pre-malaria memory B cells maintain IgM as progenies; however, about 20% only have isotype-switched progenies detected while the remaining 20% have both IgM and isotype switched progenies (**Figure 3.15c**). These pre-malaria IgM memory B cells largely retain IgM expression while further introducing SHM upon rechallenge. Thus, these analyses show multi-facet diversification potential of young children's memory B cells in a natural infection rechallenge.

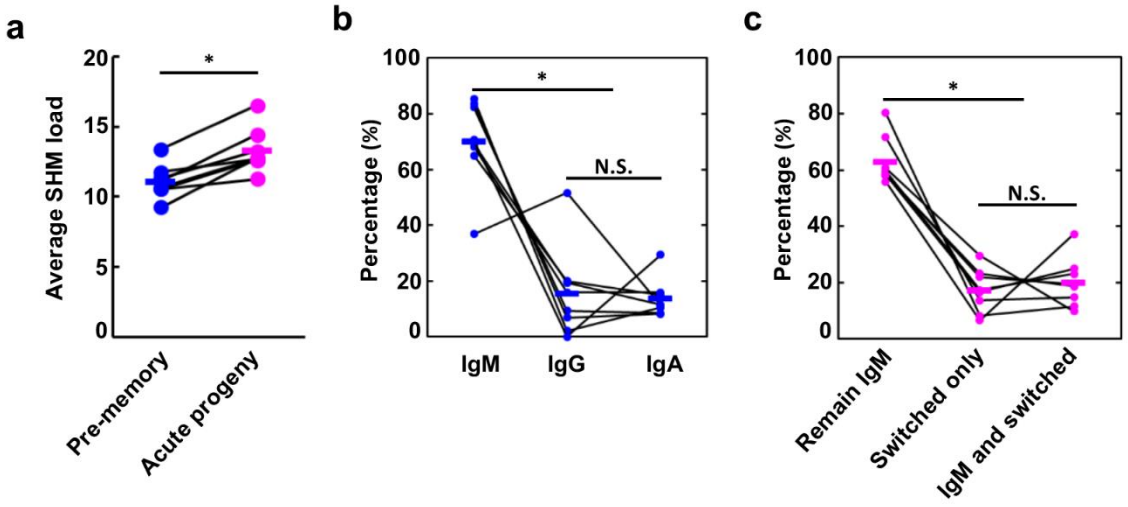

Figure 3.15: Memory B cells further mutate and class switch upon malaria rechallenge. (**a**) Average SHM load for pre-malaria memory B cells with acute progeny (blue) and their acute progenies (pink) for malaria-experienced toddlers with FACS sorted pre-malaria memory B cells (N=8). (**b**) Isotype distribution of pre-malaria memory B cells with acute progeny. (**c**) Isotype fate of acute progenies stemming from IgM pre-malaria memory B cells. Lines connect the same individuals. Bars indicate means. *$P < 0.05$, N.S. indicates not significant by two-tailed Wilcoxon Signed-Rank test.

3.3 DISCUSSION

About 13,000 children under one year old die every day worldwide[40], and most of these deaths are caused by infection[9]. It has long been recognized that children's immune systems are immature at birth and require time to develop to provide protection against pathogens or respond to vaccines. However, few studies have focused on children's antibody repertoire development, diversification, and response to infection. Knowledge in this area holds great interest to vaccine development and vaccination strategy design. This is especially urgent for malaria, as it still kills about half a million children each year[41], and the most advanced malaria vaccine confers only partial, short-lived protection in African children[42].

However, studying the antibody repertoire in young children is challenging in several regards: first, lack of analytical tools to exhaustively study the antibody repertoire from small volumes of blood; second, lack of informatic analysis tools to turn high-throughput data into knowledge; and third, the rarity of a set of samples from young children obtained before and at the time of a natural infection. To address these challenges, we developed a highly accurate and high-coverage immune repertoire sequencing tool, MIDCIRS, to analyze antibody repertoire development, diversification, and capacity to respond to a natural infection in children who were experiencing acute febrile malaria.

Previous studies showed that there was evidence of SHM and antigen selection in infants 8 months of age or older by examining a few V gene alleles[18]. However, it is not clear how widespread SHMs are in infant antibody repertoires and to what degree SHMs can be introduced in response to an infection. By using a comprehensive and unbiased analysis, here we show that infants as young as 3 months old can have 10% of sequences with 5 or more mutations, and they can further introduce mutations upon an acute febrile malaria infection to well over 20 SHM per 270nt heavy chain V region. Compared with toddlers, there is a separation on SHM load around 12 months: this number gradually increases before 12 months and stays at a plateau after that regardless of repeated malaria incidents. Consistent with this trend is the similar pattern observed in the increase in the percentage of memory B cells and corresponding decrease in percentage of naive B cells with age: both plateaued after 12 months of age. Accordantly, SHM load in IgM, IgG, and IgA correlates with the percentages of naive and memory B cells. Surprisingly, regardless of the lower mutation load in infants, their mutations are similarly, if not more strongly, selected as those of toddlers, suggesting that the molecular machineries and other cellular components involved in antibody selection are already developed in infants.

In future analyses, it will be of interest to tease out the mechanistic contributions to a two-stage increase of average mutation number, in particular the role of T cell help and germinal center formation. Regardless of these detailed mechanisms, it is clear infants can perform antibody selection as well as toddlers and adults, which provides some assurance of the effectiveness of vaccination in young children.

Another interesting finding is that infant antibody repertoires are capable of diversification, as seen through the clonal lineage analysis. We show that both infants and toddlers diversify their repertoire to the same degree upon an acute febrile malaria infection. Although adults have a larger range on the diversity of the lineage because they have accumulated more mutations, the degree of repertoire diversification in infants and toddlers is similar to what adults experience at an acute febrile malaria. This contrasts how elderly individuals preferentially increase the size of these lineages due to clonal expansion after influenza vaccination[7]. Together, these data provide evidence that infant antibody repertoires are similarly capable of responding to malaria as toddlers and adults.

Analyzing antibody lineages formed by sequences from both pre-malaria and acute malaria samples, we see an increase in SHM upon malaria infection. While the high background SHMs at pre-malaria prevented us from detecting an increase of SHM upon malaria infection in the bulk repertoire except in IgM, our two-timepoint-shared lineage analysis shows both infants and toddlers increase SHMs upon acute malaria infection. This highlights the power of combining informatics analysis with improved technology – the magnitude of SHM increase in toddlers is relatively small compared with infants, which would have been missed if MIDs were not used. Similar problems may prevent a recent study from identifying an increase of SHM in healthy adults who received influenza vaccination[21]. It is also possible that the perturbation of a natural infection is much stronger than vaccination. The evidence we observe that infants have a similar

propensity to diversify clonal lineages, perform antigen selection, and increase SHM load in acute malaria also provides a comprehensive assessment of the capacity of infants' antibody repertoire to respond to a natural infection. These results also provide some assurance that infants are capable of responding to external stimuli and develop significantly diversified antibody lineages. These, along with the observation that administering anti-parasite drugs in the first 10 months of life correlated with higher malaria incidence in second year of life[43], suggest that repeated vaccination might mimic the naturally acquired clinical malaria resistance that arises through repeated malaria exposures[12, 43].

Using progeny tracing in two-timepoint-shared lineages containing pre-malaria memory B cells, we were able to detect total of 1799 lineages that contained acute progenies. We analyzed conventional memory B cells because the atypical memory B cell frequency is low in this cohort of young children (**Figure B.8**). This is possibly due to the relatively few malaria incidents in these young children, as opposed to adults and older children that we and others have analyzed where atypical memory B cell increases prevalence as a result of many years of chronic malaria exposure[44, 45]. In depth analysis shows both continued SHM and isotype switching in memory B cells in response to a natural infection rechallenge. Although we do not know the cell phenotype of these progeny sequences, it is reasonable to speculate that some of them could have developed into plasmablasts. A previous study has shown that plasmablasts have about 30 to 35 times more transcripts per cell compared to other peripheral B cell populations[46]. Even when we relaxed the threshold and counted all unique sequences that had more than 10 copies as plasmablasts, at most 2.1% of unique sequences in these pre-malaria memory B cell-derived progenies could be inferred as plasmablasts (**Figure B.9**). Applying the same threshold to bulk PBMC sequences from acute malaria timepoint resulted in a similar

percentage of plasmablasts-derived sequences, which is consistent with about 2.2% of plasmablasts by flow cytometry analysis. All together, these data suggest that most of the progenies from pre-malaria memory B cells remain memory B cells and continued to increase SHMs in these young children.

In summary, we developed an improved method of utilizing MIDs to significantly enhance the coverage and dynamic range of IR-seq. Using it, we systemically studied the antibody repertoire in malaria-exposed infants and toddlers and discovered several aspects of repertoire development, diversification, and capacity to respond to an infection that were not known before, which provides not only new parameters and approaches in quantifying vaccine efficacy beyond traditional serological titer but also venues for future studies of detailed molecular and cellular mechanisms that drive antibody repertoire differences between infants and toddlers.

3.4 METHODS

3.4.1 Study design and cohort

Infant and toddler PBMC samples from 22 residents of Kalifabougou, Mali, ranging from 3 months old to 47 months old, were collected from an ongoing malaria cohort study[12] and analyzed as summarized in **Table B.1**. Enrollment exclusion criteria were hemoglobin level <7 g/dL, axillary temperature ≥37.5°C, acute systemic illness, use of antimalarial or immunosuppressive medications in the past 30 days, and pregnancy. The research definition of malaria was an axillary temperature of ≥37.5°C, ≥2500 asexual parasites/μL of blood, and no other cause of fever discernible by physical exam. The Ethics Committee of the Faculty of Medicine, Pharmacy, and Dentistry at the University of Sciences, Technique, and Technology of Bamako, and the Institutional Review Board of the National Institute of Allergy and Infectious Diseases, National Institutes of Health,

approved the malaria study, from which we obtained frozen PBMCs. Written informed consent was obtained from adult participants and from the parents or guardians of participating children. The study is registered in the ClinicalTrials.gov database (NCT01322581).

For this study, individuals were chosen based on the availability of frozen PBMCs in the age range specified. Blood draws were taken before the rainy season, when mosquitos are not rampant and the cases of malaria are low, and during acute febrile malaria (**Figure 3.1**). Samples were labeled for analysis by the group (Inf/Tod), patient ID, timepoint, and age at blood draw (in months), e.g. Inf1-Pre3m represents the pre-malaria timepoint for infant 1 who was 3 months old at the time of blood draw (**Table B.1**). Samples collected before the beginning of the rainy season that tested PCR negative for *Plasmodium falciparum* and *Plasmodium malariae* were designated "pre-malaria". Samples collected 7 days into acute febrile malaria infection were designated "acute malaria". Among them, 2 individuals were tracked for 2 consecutive malaria seasons and only the pre-malaria timepoint was obtained and analyzed for 9 individuals around the infant to toddler transition, as indicated in **Table B.1**. Authors were not blinded to the age group allocation or to the sample collection time.

3.4.2 Cell sorting

Frozen PBMCs were thawed, washed once in RPMI, and then processed for cell staining and sorting following standard FACS staining protocol[47, 48]. Up to 5,000,000 PBMCs were lysed directly. The remaining PBMCs were analyzed via flow cytometry; plasmablasts were gated based on the phenotype of $CD4^-CD8^-CD14^-CD56^-CD19^+CD27^{bright}CD38^{bright}$, memory B cells were gated based on the phenotype of $CD4^-CD8^-CD14^-CD56^-CD19^+CD20^+CD27^+CD38^{low}$, and naive B cells were gated based on

the phenotype of $CD4^{-}CD8^{-}CD14^{-}CD56^{-}CD19^{+}CD20^{+}CD27^{-}CD38^{low}$. Naive and memory B cells were not gated on IgD or CD21, so only conventional memory B cells were analyzed. Sorted cells were lysed in RLT Plus buffer (Qiagen) supplemented with 1% β-mercaptoethanol (Sigma). The following antibody clones were obtained from Biolegend and used as 1:25 dilution: Alexa Fluor 488 OKT3 (CD3), Alexa Fluor 700 RPA-T4 (CD4), Alexa Fluor 488 HCD14 (CD14), APC-Cy7 2H7 (CD20), PE-Cy7 O323 (CD27), APC HIT2 (CD38), Alexa Fluor 488 MEM-188 (CD56), and Brilliant Violet 605 IA6-2 (IgD). The following antibody clones were obtained from BD Biosciences: PE-CF594 RPA-T8 (CD8), Brilliant Ultraviolet 395 SJ25C1 (CD19), and Brilliant Ultraviolet 737 B-ly4 (CD21). Flow cytometry and cell sorting were performed on BD FACS Aria II.

3.4.3 Bulk antibody sequencing and reads processing

Antibody sequencing libraries were prepared and processed as described in **Chapter 2: Methods 2.4.2-4**.

3.4.4 VDJ definition and mutation counts

As described in previous work, similar methods were used to define the V, D, and J gene segments for all sequences[7, 13, 49]. From the International ImMunoGeneTics information system database (IMGT, http://www.imgt.org/textes/vquest/refseqh.html)[50], human heavy chain variable gene segment sequences (249 V-exon, 37 D-exon and 13 J-exon) were downloaded. Each unique sequence was first aligned to all 249 V gene alleles. The specific V-allele with a maximum Smith-Waterman score was then assigned. In some cases, newly identified germline alleles, defined either by TIgGER[23] or our method (below), were added to the template sequences. J-segments and D-segments were then similarly assigned. The number of mutations from germline sequence was counted

as the number of substitutions from the best aligned V and J templates as previously described[7, 13, 49]. The CDR3 was omitted due to the difficulty in determining the germline sequence. The germline sequences of V, D, and J gene segments were grouped by combining similar alleles into families using IMGT designation in VDJ correlation plots. In total, 58 V, 27 D, and 6 J families were used.

3.4.5 Novel allele detection

To address the possibility of novel germline alleles inflating the observed number of mutations, new germline alleles were assembled. See **Chapter 4 Methods 4.4.2** for detailed methodology for novel allele detection. TIgGER was used as previously reported as another method to discover novel alleles[23]. TIgGER compares the mutation rate at a specific position to the overall number of mutations for sequences within the same assigned V-gene allele. Outliers within the low mutation region suggests the existence of a novel allele, and the shape of the curve can effectively distinguish between individuals homozygous and heterozygous for the novel allele.

Our new method and TIgGER have a 90% percent overlap in newly identified alleles. Discrepancies between the two methods were treated with a conservative estimation on the number of SHM, meaning we liberally included novel alleles as part of the germline gene segments. Non-overlapping novel alleles were manually inspected, and the union of novel alleles detected by TIgGER and our method was included as part of the germline gene segments. Sequences mapped to these novel alleles were excluded from our analysis, which accounts for an average 8% of all sequences (**Table B.3**).

3.4.6 Translation from nucleotide to amino acid sequences

Nucleotide sequences were translated into amino acid sequences based on codon translation as previously described[7, 13, 49]. The unique RNA sequences were inputted to

IMGT High V quest to translate into amino acid sequences. The boundary of the CDR3 is defined by IMGT numbering for Ig and two conserved sequence markers of 'Tyr-(Tyr/Phe)-Cys' to 'Trp-Gly.' CDR3 length was determined according to these anchor residues.

3.4.7 Selection pressure

The selection pressure was evaluated via BASELINe[25]. The unique RNA molecules of PBMC, populations were inputted to BASELINe and compared with the closest IMGT germline alleles. The observed number of replacement and silent mutations were compared with the expected number of mutations for the assigned germline sequence. A selection strength as measured by the probability density function (PDF) was generated using BASELINe to indicate the direction and degree for CDR (CDR1 and 2) and FWR (FWR2, and 3) regions for each unique RNA molecule first, then combined for each individual, and then further combined for each of the two groups, infants and toddlers. The PDFs were plotted and compared between infants and toddlers (**Figure 3.8**). The associated *P* values comparing two group PDFs were calculated using a numerical integration approach that is part of the BASELINe package. Samples with 5,000,000 PBMCs were subsampled to 120,000 RNA molecules. Samples with fewer PBMCs were subsampled to proportionally fewer RNA molecules according to the PBMC number.

3.4.8 Replacement/silent mutations

According to the amino acid sequence translation results and V/D/J gene templates alignment results, we counted the number of nucleotide mutations resulting in amino acid substitutions (replacement, R) or no amino acid substitutions (silent, S) in FWR region (FWR2 and 3) and CDR region (CDR1 and 2) using the IMGT/HighV-

QUEST[51]. The number of silent and replacement mutations was averaged in each age-group (Infant and Toddler) and the ratio for silent vs. replacement mutation was calculated. The CDR3 and FWR4 were omitted due to the difficulty in determining the germline sequence. FWR1 for all sequences was also omitted because it was not covered entirely by some of the primers. Samples with 5,000,000 PBMCs were subsampled to 120,000 RNA molecules. Samples with fewer PBMCs were subsampled to proportionally fewer RNA molecules according to the PBMC number.

3.4.9 VDJ usage correlation

The correlation of VDJ usage between infants and toddlers were calculated with Pearson Correlation Coefficient as the following formula (3.1) as in previous studies[7, 13, 49]:

$$\text{corr} = \frac{\sum_{v=\{V\},d=\{D\},j=\{J\}} (X_{vdj} - \langle X \rangle)(Y_{vdj} - \langle Y \rangle)}{\sqrt{\sum_{v=\{V\},d=\{D\},j=\{J\}} (X_{vdj} - \langle X \rangle)^2 * \sum_{v=\{V\},d=\{D\},j=\{J\}} (Y_{vdj} - \langle Y \rangle)^2}} \quad (3.1)$$

vdj refers to the combination of one v allele family from 58 V gene allele families ($\{V\}$), one d allele family from 27 D gene allele families ($\{D\}$), and one j allele family from 6 J gene allele families ($\{J\}$). For the reads weighted correlation, X_{vdj} and Y_{vdj} refer to the fraction of reads assigned to the respective vdj combination for individuals X and Y, respectively. $< X >$ and $< Y >$ are the average reads across all vdj combinations. For the lineage weighted correlation, these parameters refer to the fraction of lineages for each vdj allele family combination. Samples with 5,000,000 PBMCs were subsampled to 120,000 RNA molecules. Samples with fewer PBMCs were subsampled to proportionally fewer RNA molecules according to the PBMC number. vdj

combinations that were not detected in either sample being compared were omitted from the correlation calculation.

3.4.10 Clustering sequencing into clonal lineages

Sequences with similar CDR3 are possibly progenies from the same naive B cell and can be grouped into a clonal lineage. To detect the lineage structure for the antibody repertoire, we performed single linkage clustering, using a re-parameterization of the method described previously[7, 13, 32], accounting for the larger size of the CDR3 and junction in humans as compared to zebrafish. RNA sequences with the same V and J allele assignments, the same junction length, and whose junction regions differed by no more than 10% on the nucleotide level were grouped together into a lineage. This is equivalent to a biological clone that underwent clonal expansion. In order to test the robustness of this threshold, we also tried the threshold of 95% similarity for CDR3 region[32], and it did not change the overall position of each lineage in the diversity-size plot, nor did it change the linear fits of lineage distribution generated from these two threshold (**Figure B.5**). Lineage diversity is the number of unique RNA molecules within the lineage, and lineage size is the total number of RNA molecules within the lineage. Samples with 5,000,000 PBMCs were subsampled to 120,000 RNA molecules. Samples with fewer PBMCs were subsampled to proportionally fewer RNA molecules according to the PBMC number.

3.4.11 Clonal lineage diversification

In order to discuss the clonal lineage diversification, the size and diversity, as described above, were plotted against each other for pre- and acute malaria time points for each individual (**Figure 3.9**). The linear regression visualizes the average degree of diversification relative to clonal expansion. A characteristic shift towards further

diversification of clonal lineages upon acute malaria infection was evaluated by the decrease in the slope of the linear regression for each infant and toddler. The shift was calculated by the difference between the arctangents of the slopes of the linear regressions.

3.4.12 Two-timepoint-shared lineage analysis

To test the effects of acute malaria infection on the structure of clonal lineages, RNA molecules from both the pre- and acute malaria timepoints were grouped together and subjected to clustering into clonal lineages as described above. Resulting lineages that contained sequences from both the pre-malaria and acute malaria timepoints were isolated for mutational analysis. Within these shared lineages, the average number of mutations for the pre-malaria sequences was calculated alongside the average number of mutations for the acute malaria sequences (**Figure 3.12**).

3.4.13 Lineage structure visualization

Representative lineages were selected to visualize the lineage structures and the evolution of antibody sequences. Lineage structures were generated using COLT[39] (software can be downloaded here: http://www.cs.wright.edu/~keke.chen/software/colt.zip) and validated manually. We implemented a lineage visualization tool, COLT-Viz (He et al., submitted). In short, COLT considers constraints (e.g. isotype and timepoint) along with mutational patterns to build lineage trees. The height of each node is proportional to the number of RNA molecules associated with the unique sequence (size), the color of each node relates to the number of SHMs, and the distance between nodes is proportional to the Levenshtein distance between the node sequences.

3.4.14 Pre-malaria memory B cells with acute progeny lineage analysis

To determine the fate of the pre-malaria memory B cells upon acute malaria infection, two-timepoint-shared lineages were formed as described above, and lineages containing sequences from both FACS-sorted pre-malaria memory B cells and acute malaria PBMCs were isolated for further analysis. COLT was used to generate lineage tree structures. Pre-malaria memory B cells that served as parent nodes to acute malaria sequences, as exemplified (**Figure 3.13)**, were considered "pre-malaria memory B cells with acute progeny".

3.5 REFERENCES

1. Simon AK, Hollander GA, McMichael A. Evolution of the immune system in humans from infancy to old age. *Proceedings Biological sciences* **282**, 20143085 (2015).

2. Nussbaum C*, et al.* Neutrophil and endothelial adhesive function during human fetal ontogeny. *Journal of leukocyte biology* **93**, 175-184 (2013).

3. Filias A, Theodorou GL, Mouzopoulou S, Varvarigou AA, Mantagos S, Karakantza M. Phagocytic ability of neutrophils and monocytes in neonates. *BMC pediatrics* **11**, 29 (2011).

4. Adkins B, Leclerc C, Marshall-Clarke S. Neonatal adaptive immunity comes of age. *Nature reviews Immunology* **4**, 553-564 (2004).

5. Rechavi E*, et al.* Timely and spatially regulated maturation of B and T cell repertoire during human fetal development. *Science translational medicine* **7**, 276ra225 (2015).

6. Prabakaran P*, et al.* Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* **64**, 337-350 (2012).

7. Jiang N*, et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med* **5**, 171ra119 (2013).

8. Wu YC, Kipling D, Dunn-Walters DK. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front Immunol* **3**, 193 (2012).

9. PrabhuDas M*, et al.* Challenges in infant immunity: implications for responses to infection and vaccines. *Nature immunology* **12**, 189-194 (2011).

10. Portugal S*, et al.* Malaria-associated atypical memory B cells exhibit markedly reduced B cell receptor signaling and effector function. *eLife* **4**, (2015).

11. Zinocker S*, et al.* The V gene repertoires of classical and atypical memory B cells in malaria-susceptible West African children. *IEEE Trans Vis Comput Graphics* **194**, 929-939 (2015).

12. Tran TM*, et al.* An intensive longitudinal cohort study of Malian children and adults reveals no evidence of acquired immunity to Plasmodium falciparum infection. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **57**, 40-47 (2013).

13. Jiang N, Weinstein JA, Penland L, White RA, 3rd, Fisher DS, Quake SR. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5348-5353 (2011).

14. Boyd SD*, et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine* **1**, 12ra23 (2009).

15. Glanville J*, et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 20216-20221 (2009).

16. Schroeder HW, Jr., Zhang L, Philips JB, 3rd. Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood* **98**, 2745-2751 (2001).

17. Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. *Nature* **354**, 389-392 (1991).

18. Ridings J, Dinan L, Williams R, Roberton D, Zola H. Somatic mutation of immunoglobulin V(H)6 genes in human infants. *Clinical and experimental immunology* **114**, 33-39 (1998).

19. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 13463-13468 (2013).

20. Ridings J, Nicholson IC, Goldsworthy W, Haslam R, Roberton DM, Zola H. Somatic hypermutation of immunoglobulin genes in human neonates. *Clinical and experimental immunology* **108**, 366-374 (1997).

21. Ellebedy AH*, et al.* Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat Immunol* **17**, 1226-1234 (2016).

22. Biswas S, Saxena QB, Roy A, Kabilan L. Naturally occurring plasmodium-specific IgA antibody in humans from a malaria endemic area. *Journal of Biosciences* **20**, 453-460 (1995).

23. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E862-870 (2015).

24. Chang B, Casali P. The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. *Immunology today* **15**, 367-373 (1994).

25. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic acids research* **40**, e134 (2012).

26. O'Brien PM, Tsirimonaki E, Coomber DW, Millan DW, Davis JA, Campo MS. Immunoglobulin genes expressed by B-lymphocytes infiltrating cervical carcinomas show evidence of antigen-driven selection. *Cancer immunology, immunotherapy : CII* **50**, 523-532 (2001).

27. Machida K*, et al.* Hepatitis C virus induces a mutator phenotype: enhanced mutations of immunoglobulin and protooncogenes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4262-4267 (2004).

28. Schroder AE, Greiner A, Seyfert C, Berek C. Differentiation of B cells in the nonlymphoid tissue of the synovial membrane of patients with rheumatoid arthritis. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 221-225 (1996).

29. Weitkamp JH, Lafleur BJ, Greenberg HB, Crowe JE, Jr. Natural evolution of a human virus-specific antibody gene repertoire by somatic hypermutation requires both hotspot-directed and randomly-directed processes. *Human immunology* **66**, 666-676 (2005).

30. Haynes BF, Kelsoe G, Harrison SC, Kepler TB. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nature biotechnology* **30**, 423-433 (2012).

31. Doria-Rose NA*, et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509**, 55-62 (2014).

32. Horns F*, et al.* Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *eLife* **5**, (2016).

33. Dogan I*, et al.* Multiple layers of B cell memory with different effector functions. *Nat Immunol* **10**, 1292-1299 (2009).

34. Pape KA, Taylor JJ, Maul RW, Gearhart PJ, Jenkins MK. Different B cell populations mediate early and late memory during an endogenous immune response. *Science* **331**, 1203-1207 (2011).

35. Kaji T*, et al.* Distinct cellular pathways select germline-encoded and somatically mutated antibodies into immunological memory. *J Exp Med* **209**, 2079-2097 (2012).

36. Weisel FJ, Zuccarino-Catania GV, Chikina M, Shlomchik MJ. A Temporal Switch in the Germinal Center Determines Differential Output of Memory B and Plasma Cells. *Immunity* **44**, 116-130 (2016).

37. Krishnamurty AT*, et al.* Somatically Hypermutated Plasmodium-Specific IgM(+) Memory B Cells Are Rapid, Plastic, Early Responders upon Malaria Rechallenge. *Immunity* **45**, 402-414 (2016).

38. Taylor JJ, Jenkins MK, Pape KA. Heterogeneity in the differentiation and function of memory B cells. *Trends in immunology* **33**, 590-597 (2012).

39. Chen K, Gogu V, Wu D, Jiang N. COLT: Constrained Lineage Tree Generation from Sequence Data. *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, (2016).

40. UNICEF. The State of the World's Children 2015: Reimagine the Future: Innovation for every child. *United Nations Children's Fund, New York, 2015*, (2015).

41. WHO. *World Malaria Report 2015*. Global Malaria Programme, World Health Organization (2015).

42. White MT*, et al.* A combined analysis of immunogenicity, antibody kinetics and vaccine efficacy from phase 2 trials of the RTS,S malaria vaccine. *BMC medicine* **12**, 117 (2014).

43. Guinovart C*, et al.* The role of age and exposure to Plasmodium falciparum in the rate of acquisition of naturally acquired immunity: a randomized controlled trial. *PLoS One* **7**, e32362 (2012).

44. Weiss GE*, et al.* A positive correlation between atypical memory B cells and Plasmodium falciparum transmission intensity in cross-sectional studies in Peru and Mali. *PLoS One* **6**, e15983 (2011).

45. Weiss GE*, et al.* The Plasmodium falciparum-specific human memory B cell compartment expands gradually with repeated malaria infections. *PLoS Pathog* **6**, e1000912 (2010).

46. Shi W*, et al.* Transcriptional profiling of mouse B cell terminal differentiation defines a signature for antibody-secreting plasma cells. *Nat Immunol* **16**, 663-673 (2015).

47. Yu W*, et al.* Clonal Deletion Prunes but Does Not Eliminate Self-Specific alphabeta CD8(+) T Lymphocytes. *Immunity* **42**, 929-941 (2015).

48. Zhang SQ*, et al.* Direct measurement of T cell receptor affinity and sequence from naive antiviral T cells. *Sci Transl Med* **8**, 341ra377 (2016).

49. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807-810 (2009).

50. Lefranc MP*, et al.* IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* **43**, D413-422 (2015).

51. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* **882**, 569-604 (2012).

Chapter 4: A streamlined approach to antibody novel germline allele prediction and validation[3]

4.1 INTRODUCTION

V(D)J recombination and non-template nucleotide insertion in the junction regions generate the first level of antibody repertoire diversity. During an immune response, B cells that are activated by binding their matching antigens go through a clonal expansion process accompanied by somatic hypermutations (SHMs) that are quasi-randomly introduced to the antibody genes. These mutated antibodies are then selected based on binding strength to the antigen, leading to a second generation of higher affinity antibodies[1, 2, 3].

Antibody repertoire SHM patterns have been implicated in a wide range of applications, from the development of broadly neutralizing antibodies against HIV to the diminished effectiveness of vaccines in elderly subjects[4, 5, 6]. The recent incorporation of molecular barcodes into high-throughput immune repertoire sequencing has improved the ability to discern individual SHMs from PCR and sequencing errors[7, 8]; however, accurate SHM calling requires an accurate set of reference germline sequences to align to. The polygenic and polyallelic nature of the variable domain locus confounds this issue. Currently, 259 functional human antibody heavy chain V gene alleles listed in the IMGT database can be broken into 7 subfamilies that likely share common evolutionary ancestors based on sequence similarity[9], but recent studies have shown that individuals often carry novel alleles that have yet to be characterized in the IMGT database[10, 11, 12].

---

[3]Wendel, *et al*. A streamlined approach to antibody novel germline allele prediction and validation. *Frontiers in Immunology*, under review. B.S.W. designed and performed research, analyzed and interpreted data, and wrote the manuscript. C.H. helped perform data analysis. N.J. designed research, directed the study, provided funding, and wrote the manuscript.

These novel alleles can be problematic for antibody repertoire analysis because single nucleotide polymorphisms (SNPs) between the novel alleles and the nearest IMGT alleles will instead be counted as SHMs on every sequence utilizing that allele, inflating the SHM load and skewing the SHM patterns. Although there are several software tools[10, 11] to predict the existence of novel alleles, a simple method for novel allele prediction and validation is lacking, especially using a small amount of blood samples.

Here, we report a streamlined method for predicting novel alleles from bulk antibody repertoire data and validating them by sequencing unrecombined genomic DNA (gDNA) from FACS-sorted non-B cells. This method can be applied to PBMCs and B cells purified from as little as 4ml of blood. 6 novel alleles across 8 different subjects from a larger, ongoing malaria study cohort[13] were predicted and validated with perfect congruency between the expressed repertoire and gDNA. This method can quickly and easily be applied to any antibody repertoire data to mitigate the effects of germline mismatches on SHM patterns.

4.2 RESULTS

4.2.1 Bulk antibody repertoire novel allele prediction

Antibody repertoire sequencing data from bulk PBMCs were collected and processed as described in **Methods 4.4.2** and summarized in **Figure 4.1**. IgM sequences were used to calculate the mutation distribution by position because IgM is mostly expressed on naïve B cells that have not been activated and have fewer SHMs compared to other isotypes. As expected, the percentage of unique sequences mutated at each position in IgM is low, even for SHM hotspots (**Figure 4.2a**). However, a large spike at one (or more) specific position(s) could indicate the presence of a novel allele resulted from SNP(s) (**Figure 4.2b**).

A threshold of 20% of unique sequences harboring the identical predicted SNP(s) was applied to determine a positive hit on novel allele (**Figure 4.2**, dashed horizontal line). Several *IGHV* genes, e.g. *IGHV1-69* and *IGHV3-30*, have copy-number variants (CNVs) that arose from chromosomal segmental duplication and insertion/deletion events, leading to a diploid copy number ranging from 0 to 4 alleles present for a given gene for an individual[14, 15]. For genes with up to 4 copies, this 20% threshold can account for a heterozygous genotype with 1 of 4 copies being the novel allele, which should roughly have a 25:75 split on the usage of these four alleles. 6 genes with predicted novel alleles were chosen for validation (**Table 4.1**, column headers). These novel alleles were also predicted independently using TIgGER[10], another novel germline allele detection tool. Overall, 17 positive novel allele hits were predicted from the 6 genes across the 8 subjects.

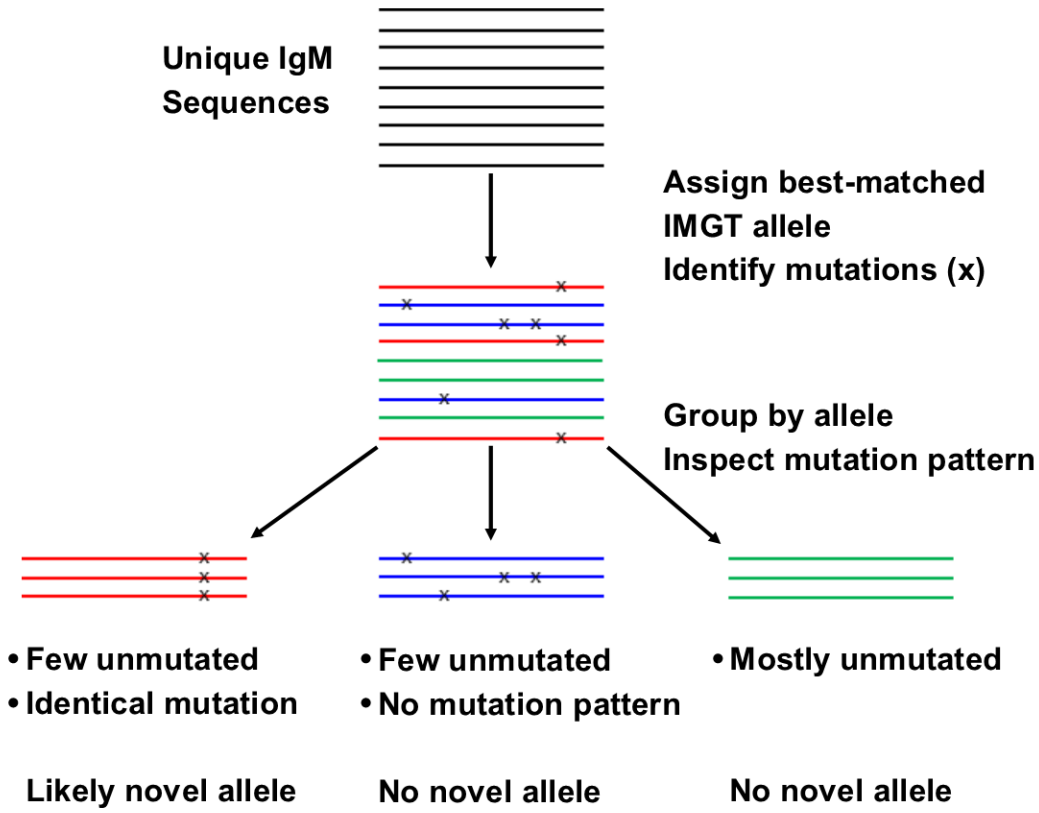

Figure 4.1: Overview of novel allele prediction schematic from bulk repertoire sequencing data. Color indicates best-matched IMGT reference germline allele assignment; x indicates SNP to reference germline allele.

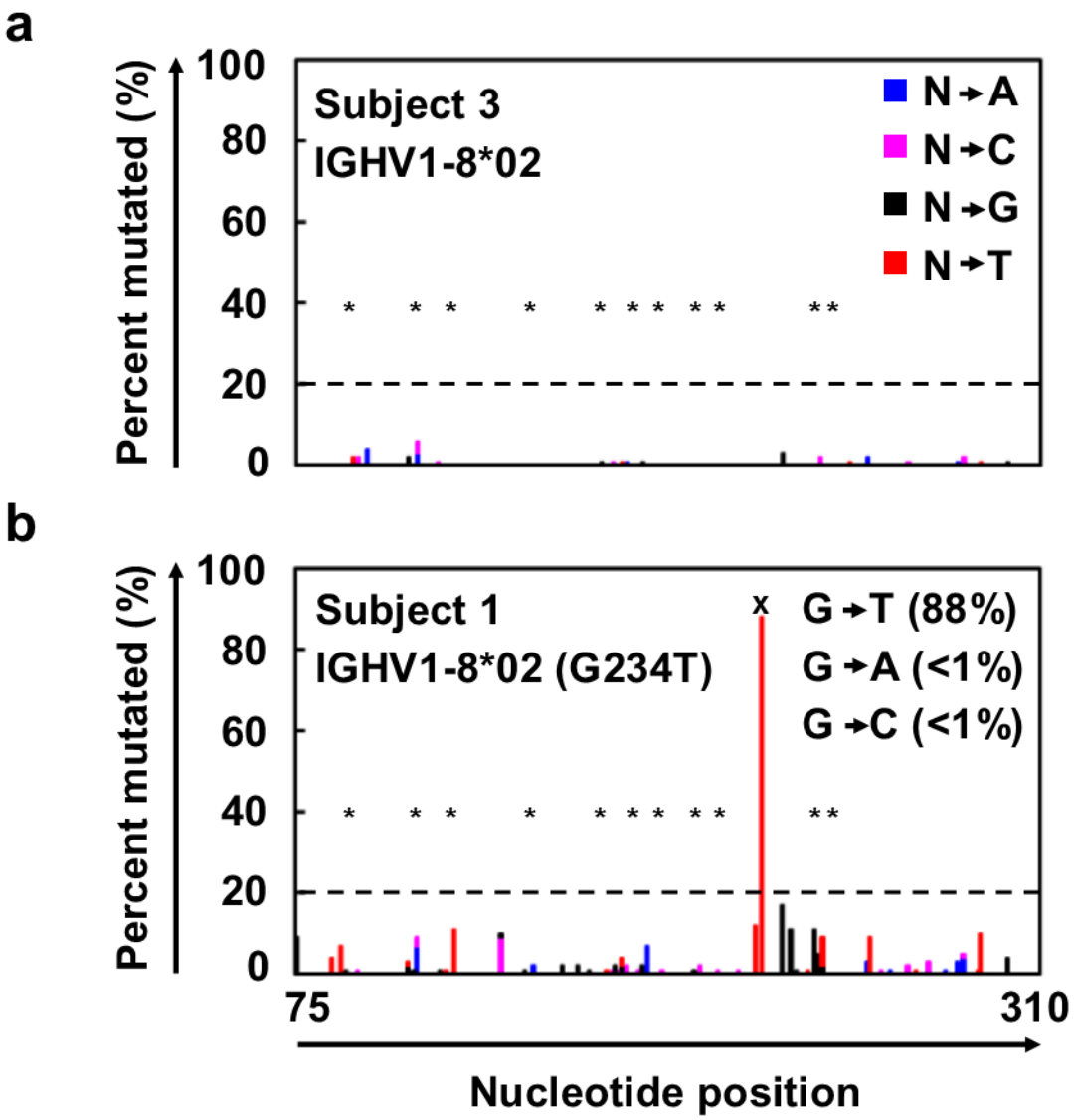

Figure 4.2: Novel germline allele prediction from bulk repertoire sequencing data. Representative percent of unique IgM sequences mutated for each position along the sequence for the absence (**a**) and presence (**b**) of a novel allele resulted from SNP(s). Color indicates the nucleotide substitution: A (blue), C (pink), G (black), and T (red). * indicates SHM hotspots; dashed line indicates prediction threshold of SNP calling; x indicates SNP on predicted novel allele compared to the closest IMGT reference germline allele. SNP is broken down by nucleotide substitution as indicated in inset in (**b**).

Subjects	Novel alleles					
	IGHV1-8*02 (G234T)	IGHV1-69*01 (G163A)	IGHV3-30*02 (T201C)	IGHV4-31*02 (C198T)	IGHV4-59*01 (T109C)	IGHV4-61*01 (C93T_C136G_A138C)
1	+/+	+/+	+/+*	-/-	+/+	-/-
2	-/N.D.	-/-	-/-	+/+	-/-	-/-$
3	-/-	+/+*	-/-	-/-	+/+	+/+
4	+/+	-/-	-/-*	-/-	-/-	-/-
5	-/-	-/-	-/-	+/+	-/-	-/-
6	+/+	-/-*	+/+	-/-$	+/+	-/-$
7	+/+	-/-	+/+*	-/-$	-/-	-/-
8	+/+	+/+	-/-	-/-$	-/-	-/-

Table 4.1: Summary of gDNA validation of novel alleles predicted by bulk repertoire sequencing data. +/+ (dark green) indicates positive in both the bulk repertoire and the gDNA data for predicted SNPs; -/- (light green) indicates negative in both the bulk repertoire and the gDNA data for predicted SNPs; -/N.D. (yellow) indicates negative in bulk repertoire data but gDNA failed to amplify during gDNA validation for predicted SNPs. * indicates the existence of CNVs with more than 2 alleles detected in the gDNA data that belong to the same gene. $ indicates the gene was not detected in the repertoire or gDNA, possibly due to gene deletion.

4.2.2 gDNA novel allele validation

gDNA purified from FACS-sorted T cells from the same 8 subjects was used to validate the presence of the predicted novel alleles as described in **Methods 4.4.3** and summarized in **Figure 4.3**. Due to the high degree of sequence homology among the V genes, a series of filtering steps was applied to eliminate reads that were distant from the putative novel allele or closest IMGT allele. Finally, the number of reads exactly matching the novel and original IMGT sequences were compared. If 20% of these reads matched the novel allele sequence, the subject was deemed positive for the novel allele. All 17 of the positive hits from the bulk repertoire data returned positive hits from the gDNA, and 30 out of 31 negative hits from the bulk repertoire data that were tested in parallel were also negative in the gDNA, with one library failing to amplify (**Table 4.1**).

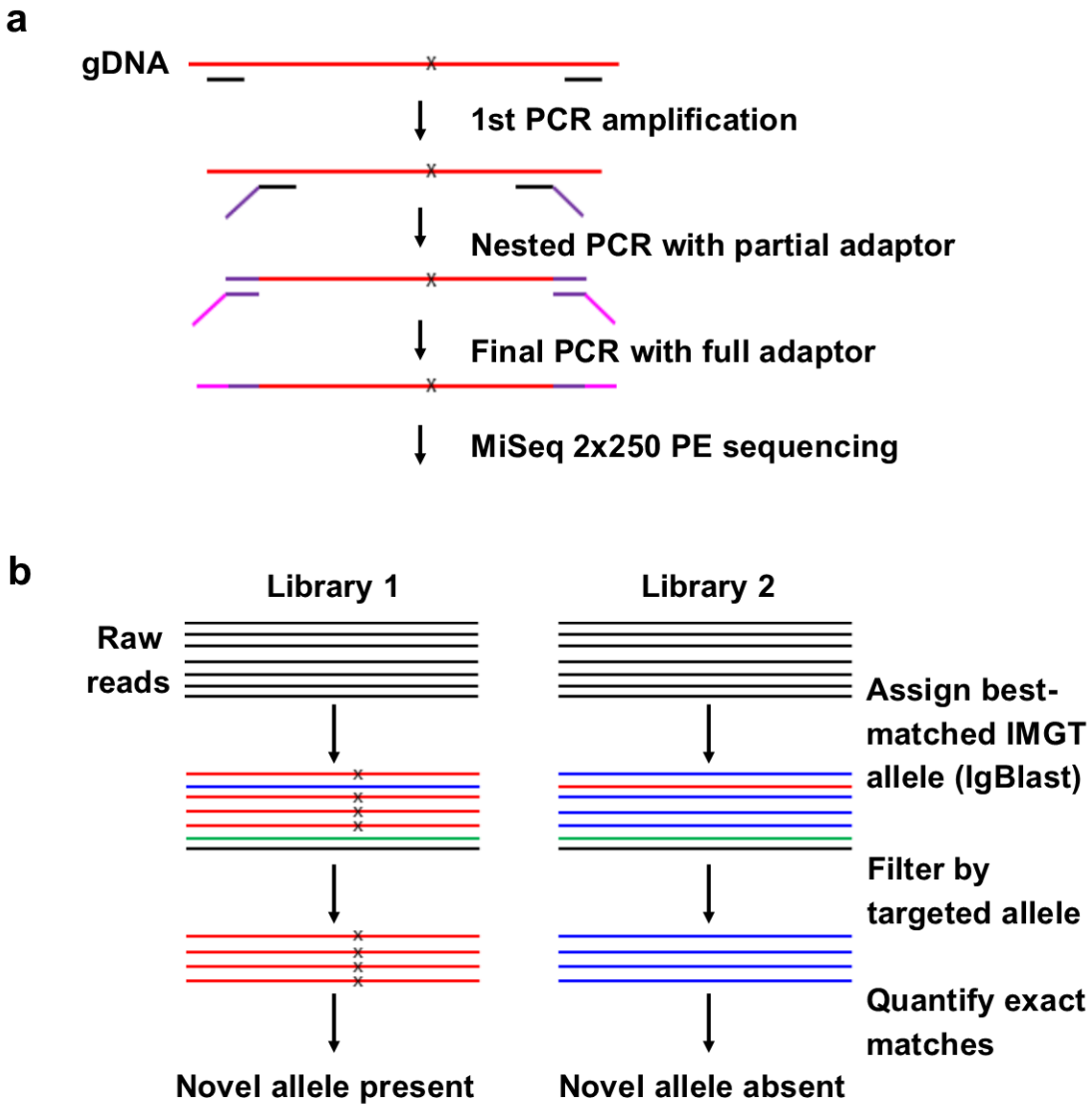

Figure 4.3: Novel allele validation by targeted gDNA sequencing. (**a**) Overview of targeted gDNA amplification and library preparation. x indicates predicted SNP on novel allele compared to the closest IMGT reference germline allele. (**b**) Overview of gDNA sequencing data analysis for the presence (left) and absence (right) of a novel allele resulted from SNP(s). Color indicates best-matched IMGT reference germline allele assignment; x indicates SNP to the closest IMGT reference germline allele.

The positive hits in the bulk repertoire data ranged from 22.7% - 99.9% of unique sequences containing the novel mutation, while the negative hits ranged from 0.0% - 1.6% (**Figure 4.4**, x-axis). This tight range on the negative hits is consistent with the low rate of mutations expected for IgM antibodies. For the gDNA validation, the positive hits ranged from 29.0% - 100% of filtered reads exactly matching the novel sequence, while the negative hits all failed to detect a single novel allele read (**Figure 4.4**, y-axis). The densely packed clusters at the bottom left and top right of **Figure 4.4** imply that this method is sensitive enough to distinguish between heterozygous and homozygous genotypes, and our threshold of calling a novel allele on both gDNA and repertoire data, which is 20% of reads mapped to either putative allele or its closest IMGT allele, is appropriate.

Another observation that increases confidence in novel germline allele prediction is the detection of the identical novel allele in multiple individuals[11]. 5 of the 6 alleles tested were positively validated in 2 or more subjects (**Table 4.1**). The lone allele detected in a single individual, *IGHV4-61*01 (C93T_C136G_A138C)*, is 3 mismatches away from the nearest IMGT germline allele. None of the gDNA reads for this individual matched the reference allele, while all of the filtered reads exactly matched the predicted novel sequence. Additionally, none of the filtered reads from all 7 negative subjects tested in parallel matched the novel sequence.

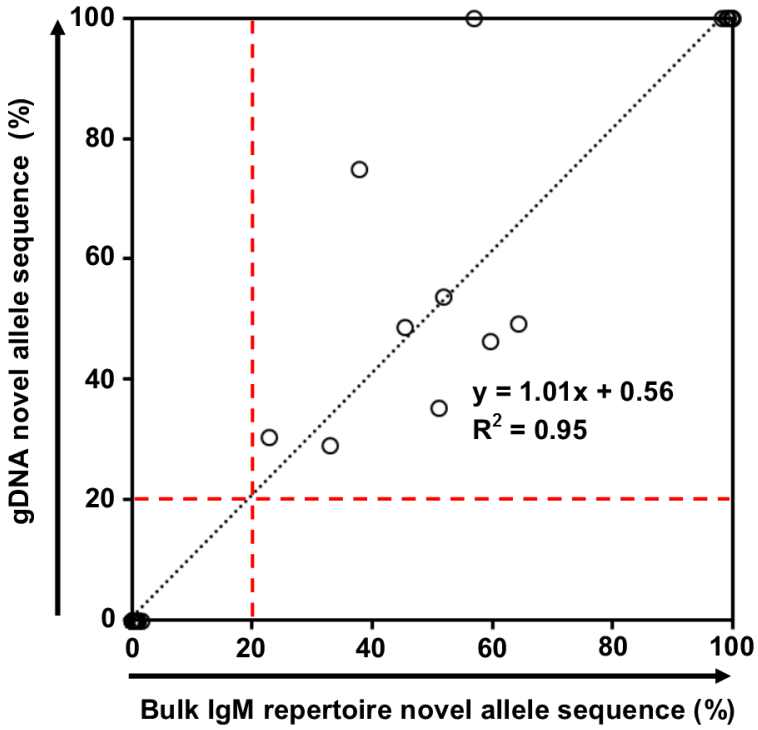

Figure 4.4: Novel germline allele prediction and validation congruency. Correlation between the percentage of novel allele sequences in bulk IgM repertoire data (%) and the percentage of novel allele sequences in gDNA data (%). Most points are clustered at the origin (N = 30) or the top right (N = 9). Black dotted line represents the linear regression; red dashed lines indicate the novel allele calling threshold.

4.3 DISCUSSION

We developed a remarkably sensitive yet simple method for detecting and validating novel alleles from bulk antibody repertoire sequencing data. This approach requires little specialized bioinformatics analysis and no unique laboratory equipment or reagents. gDNA validation can be performed on DNA purified from as few as 2,000 FACS-sorted non-B cells, maximizing the proportion of the sample that can be utilized for antibody repertoire analysis.

The combination of antibody repertoire and non-B cell gDNA sequencing allowed for advanced insight into the genotypes of the subjects. Novel allele *IGHV1-69*01*

(G163A) was found to be an exact match with IMGT allele *IGHV1-69*07*, except *IGHV1-69*07* is truncated at both ends. After performing nested PCR, these alleles were indistinguishable. However, in the 3 subjects predicted to have the novel allele, no unique sequences in the bulk repertoire data mapped to the truncated *IGHV1-69*07* allele; instead, they contained the full length *IGHV1-69*01* sequence with the G163A SNP.

IGHV1-69 and *IGHV1-69D* share common alleles that can range from 2 to 4 copies total on a diploid genome, and *IGHV3-30* and *IGHV3-30-5* share common alleles that can range from 0 to 4 copies total on a diploid genome[15]. Interestingly, we detected more than 2 alleles in the gDNA of 2 of 8 subjects for *IGHV1-69/1-69D*, consistent with previous studies on *IGHV1-69* CNV in African populations[14], and more than 2 *IGHV3-30/3-30-5* alleles were detected in 3 of 8 subjects (* in **Table 4.1**). Conversely, *IGHV4-31* and *IGHV4-61* are associated with deletion events yielding 0 to 4 copies, each[15]. *IGHV4-31* was not observed in the repertoire or gDNA of 3 of 8 subjects, and *IGHV4-61* was not observed in the repertoire or gDNA of 2 of 8 subjects ($ in **Table 4.1**), likely indicating the absence of these genes in these subjects. These results demonstrated the sensitivity of our approach and emphasized the necessity of characterizing individual's own germline alleles in antibody repertoire sequencing studies in order to accurately count number of SHMs.

The results were highly consistent with all 17 predicted positive hits and 30 of 31 predicted negative hits confirmed in gDNA. One limitation is that this method will only detect novel alleles that are similar to alleles within the IMGT reference database. Additionally, if a CNV results in more than 4 alleles present in the diploid genome for a given gene in an individual, then our threshold of a putative SNP call, which is least 20% of unique IgM sequences having the same mismatch at the same position, would not be able to detect the novel allele initially in the antibody repertoire data. However, this is

extremely rare based on current knowledge of antibody gene loci[15]. In summary, at least 1 novel allele was found in each subject tested, highlighting the need for novel allele detection and correction in antibody repertoire analysis.

4.4 METHODS

4.4.1 Study design and cohort

PBMC samples from 8 residents of Kalifabougou, Mali were collected from an ongoing malaria cohort study[13]. Up to 5 million PBMCs were directly lysed for antibody repertoire sequencing, and T cells were FACS-sorted from the remaining PBMCs for unrecombined gDNA validation. 6 predicted novel alleles were chosen for validation. The Ethics Committee of the Faculty of Medicine, Pharmacy, and Dentistry at the University of Sciences, Technique, and Technology of Bamako, and the Institutional Review Board of the National Institute of Allergy and Infectious Diseases, National Institutes of Health, approved the malaria study, from which we obtained frozen PBMCs. Written informed consent was obtained from adult participants and from the parents or guardians of participating children. The study is registered in the ClinicalTrials.gov database (NCT01322581).

4.4.2 Antibody repertoire sequencing and novel allele prediction

Antibody repertoire sequencing was performed as described in **Chapter 2 Methods 2.4.2-4**. Novel allele prediction schematic is summarized in **Figure 4.1**. In short, unique IgM sequences from bulk PBMC samples were used to minimize the effects of SHM and clonal expansion, as they are more likely to be derived from naïve B cells and thus have fewer SHMs than other antibody isotypes. These sequences were first aligned to the reference germline allele database (e.g. IMGT) and assigned to the best-matched alleles. The ratios of perfectly matched sequences to those with 1, 2, 3 and 4

mismatches (putative SNPs) compared to the reference germline were determined. Ratios of less than 2 to 1 were then inspected for identical mutation patterns. If identical mutations were present in at least 20% of the unique sequences, with less than 2% of the sequences harboring different mutations at the same positions, the allele containing those SNPs was flagged as a possible novel allele.

4.4.3 gDNA sequencing and reads processing

Nested PCR was used to reduce nonspecific amplification. Primers were designed such that the inner primers were no fewer than 14 bases away from the locations of the predicted IMGT/novel allele mismatches. Inner primers were fused to partial Illumina adaptors, and a third PCR was performed to add the full adaptor sequence (**Table C.1**). 1st PCR was performed on 10% of purified gDNA from 2,000 sorted T cells using Phusion Hot Start II DNA Polymerase (Thermo Scientific) with the following protocol: 98 ^{0}C for 1 minute; 10 cycles of 98 ^{0}C for 30 seconds, 57 ^{0}C for 1 minute, and 72 ^{0}C for 5 minutes; then 72 ^{0}C for 10 minutes. 2nd PCR was performed on 10% of the 1st PCR product with the same protocol. Final adaptor ligation was performed on 10% of the 2nd PCR product using TaKaRa Ex Taq DNA Polymerase Hot Start with the following protocol: 95 ^{0}C for 3 minutes; 10 cycles of 95 ^{0}C for 30 seconds, 57 ^{0}C for 30 seconds, and 72 ^{0}C for 2 minutes; then 72 ^{0}C for 7 minutes. Libraries were pooled, gel-purified, and sequenced via Miseq 2x250 PE.

Sequencing reads were first merged using the SeqPrep tool ((https://github.com/jstjohn/SeqPrep). IgBlast[16] was then used to align the reads to the established IMGT germline allele database. Reads mapping to the nearest germline allele to the novel allele of interested were filtered. Reads matching exactly to the IMGT germline allele or the novel allele sequence were tallied. If the exact novel allele

sequence was found in 20% or more of the tallied reads, the sample was considered a positive hit.

4.5 REFERENCES

1. Victora GD, Nussenzweig MC. Germinal centers. *Annual review of immunology* **30**, 429-457 (2012).

2. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. *Annual review of biochemistry* **76**, 1-22 (2007).

3. De Silva NS, Klein U. Dynamics of B cells in germinal centres. *Nature reviews Immunology* **15**, 137-148 (2015).

4. Jiang N*, et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine* **5**, 171ra119 (2013).

5. Zhu J*, et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 6470-6475 (2013).

6. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Current opinion in immunology* **25**, 646-652 (2013).

7. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 13463-13468 (2013).

8. Shugay M*, et al.* Towards error-free profiling of immune repertoires. *Nature methods*, (2014).

9. Lefranc MP*, et al.* IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic acids research* **43**, D413-422 (2015).

10. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci U S A* **112**, E862-870 (2015).

11. Corcoran MM*, et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications* **7**, 13642 (2016).

12. Boyd SD*, et al.* Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* **184**, 6986-6992 (2010).

13. Tran TM*, et al.* An intensive longitudinal cohort study of Malian children and adults reveals no evidence of acquired immunity to Plasmodium falciparum infection. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **57**, 40-47 (2013).

14. Watson CT*, et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *American journal of human genetics* **92**, 530-546 (2013).

15. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes and immunity* **13**, 363-373 (2012).

16. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research* **41**, W34-40 (2013).

Chapter 5: HIV-driven T cell expansion promotes restricted functional diversity in germinal center follicular helper T cells[4]

5.1 INTRODUCTION

Follicular helper T cells (T_{FH}) are a subset of $CD4^+$ T cells within secondary lymphoid organs that provide key signals necessary for B cell survival, maturation, and antibody production[1, 2]. In particular, CD57 expression on T_{FH} cells marks a T_{FH} cell subset that localizes to the germinal center (GC) and is required for isotype switching, somatic hypermutation, and the selection of high affinity B cell clones[3, 4]. During chronic HIV infection, GC T_{FH} cells become a primary target of HIV infection. Studies of lymph nodes (LN) from untreated HIV^+ patients have detected substantial enrichment of p24 antigen expression predominantly in GC and up to 10-fold higher frequency of HIV proviral DNA containing cells within the $CD57^+$ subset compared to other $CD4^+$ T cells[5, 6]. Thus, GC T_{FH} cells play a crucial role in lymphoid biology and are also a major HIV reservoir contributing to viral persistence during chronic HIV infection.

Paradoxically, while the frequency of T_{FH} cells and serum immunoglobulin become elevated during chronic HIV infection, HIV^+ patients develop a less efficient antibody response to immunization[7, 8]. For example, HIV-infected patients produce lower titers of antibodies and less durable responses to seasonal influenza vaccines[9, 10, 11]. Cubas et al. used T_{FH} cells from HIV^+ LNs to show that PD-L1 expression on B cells dampens T_{FH} cell function[12]. We hypothesize that GC T_{FH} cells are themselves targeted

[4]Wendel, *et al*. HIV-driven T cell expansion promotes restricted functional diversity in germinal center follicular helper T cells. *Science Immunology*, under review. B.S.W. performed TCR sequencing and data analysis, C.H. analyzed TCR sequencing data, D.A.A. performed CyTOF staining and analysis. B.A. performed in vitro peptide stimulation, P.D.R. and G.R.T established the infrastructure for HIV^+ patient recruitment and provided HIV^+ LN samples and the associated clinical information. S.M.H. assisted with TCR sequencing library preparation. K.-Y.M. adapted the TCR repertoire sequencing protocol. M.B. contributed to the lymph nodes from healthy individuals. L.F.S. and N.J. designed the study. L.F.S., N.J., and B.S.W. wrote and edited the paper.

by HIV-mediated changes and acquire new functional and phenotypic characteristics. In addition, as antigen-dependence of GC T_{FH} cell expansion in HIV^+ LNs has never been evaluated, it remains unclear whether the increase in GC T_{FH} cells results from an antigen-driven process or reflects general immune activation during chronic HIV infection.

Recent breakthroughs in single cell-based protein expression analysis have led to a better appreciation of the complexity of cell populations in different tissue compartments. At the forefront of these new analysis tools is mass cytometry (CyTOF), which combines the advantages of flow cytometry with mass spectrometry to measure protein expression by the detection of heavy metal isotope-specific signatures produced by cells stained with metal ion-conjugated antibodies[13]. Here, we applied CyTOF to analyze the cellular diversity of GC T_{FH} cells and combined this approach with T cell receptor (TCR) repertoire sequencing to trace the clonal composition of GC T_{FH} cells during chronic viral infection. We hypothesize that chronic viral stimulation drives a reduction in T_{FH} cell diversity via preferential expansion of an activated and functionally distinct subset of GC T_{FH} cells. Using a unique set of LN samples obtained from HIV^+ patients, healthy controls (HCs), and patients with active inflammatory bowel disease (IBD), our data revealed HIV-specific skewing in the functional and activation states of GC T_{FH} cells. We also found evidence of antigen-driven convergent selection and expansion of these GC T_{FH} cells under chronic viral stimulation. These data provide an in-depth analysis of HIV-induced perturbation on GC T_{FH} cells and emphasize a key role for HIV infection in driving GC T_{FH} cell pathology.

5.2 RESULTS

5.2.1 GC T_{FH} cells are elevated in HIV^+ patients and acquire distinct characteristics during chronic inflammation

To evaluate the cellular diversity of GC T_{FH} cells, we designed a mass cytometry panel to examine phenotypic and functional features of T_{FH} cells (**Table D.1**). LN samples were obtained from 25 HIV^+ patients, including 20 who were treatment naïve (**Table D.2**). For comparison, we included 7 LN samples from healthy donors and, as a control for a non-HIV inflammatory disease, 4 LN samples from patients undergoing bowel resection from IBD. Cryopreserved LN cells were thawed and 3-5 million cells were stained with metal-conjugated antibodies for CyTOF. Approximately 2 - 3 million thawed LN cells were separately cultured with an HIV-1 consensus B gag peptide pool or an overlapping set of hemagglutinin (HA) peptides from influenza virus (A/California/7/2009) for 3-4 weeks to expand HIV- or influenza-specific T cells, respectively. In parallel, we sorted 1,464 to 15,000 naïve, memory, or GC T_{FH} cells using freshly thawed LN samples for TCR sequencing analysis (**Figure 5.1** and **Table D.3**).

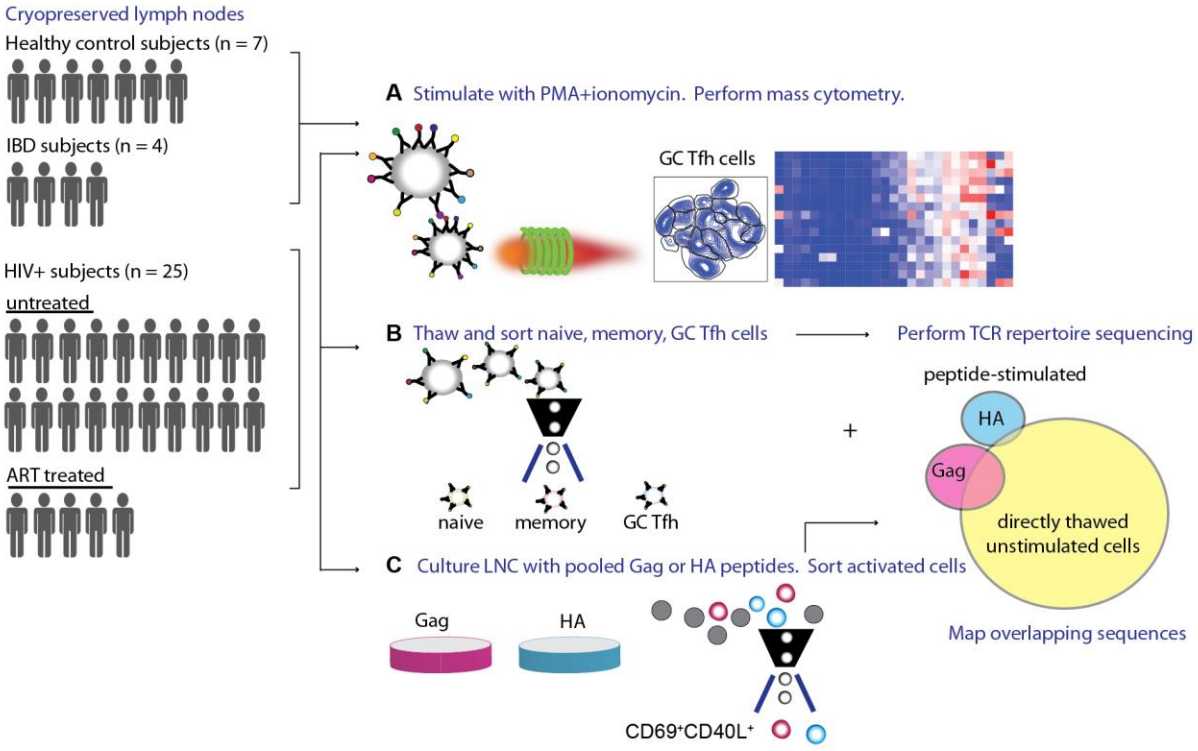

Figure 5.1: Summary of experimental design. Cryopreserved LN samples were obtained from 7 healthy controls, 4 IBD patients, and 25 HIV^+ individuals. (**A**) Cells from all donors were stimulated with PMA plus ionomycin and analyzed on CyTOF. (**B**) LN cells from 8 HIV^+ donors were sorted by naïve ($CD45RO^-CXCR5^-CCR7^+CD27^+$), memory ($CD45RO^+CXCR5^-PD1^-ICOS^-$), or GC T_{FH} phenotype ($CD45RO^+CXCR5^+PD1^+CD57^+$) for TCR sequencing. (**C**) A subset of these donors also had enough cells for peptide stimulation in culture (5 for Gag and 6 for HA). After 3-4 weeks, cultured cells were restimulated with peptides and sorted for activation by CD69 and CD40L expression. TCR sequences obtained from Gag or HA-peptide reactive T cells were used as a reference sequence dataset to identify matching HA- or Gag-specific T cells from sorted and sequenced bulk populations from (**B**). Figure produced by L.F.S.

To interrogate the functional features of GC T_{FH} cells, LN cells were stimulated with phorbol-12-myristate-13-acetate (PMA) and ionomycin in the presence of Brefeldin A and monensin for 5 hours to capture cytokine secretion. Cells were then stained with a panel of metal-conjugated antibodies and acquired on the mass cytometer CyTOF 2.

Data from each sample were normalized using a bead-based normalizer. We initially performed manual gating to identify GC T_{FH} cells (**Figure D.1**). Consistent with prior studies, our data demonstrated an increase in GC T_{FH} cells during chronic HIV infection both by relative frequency and total cell count (**Figure 5.2A,B**). To begin to interrogate the phenotypic diversity of GC T_{FH} cells, data for $CD4^+$ T cells from all donors were downsampled to obtain equal numbers of cells per disease category and imported into Cytobank for t-SNE (t-Distributed Stochastic Neighbor Embedding) using the viSNE implementation. This dimensionality reduction method creates a two-dimensional plot that places cells with similar phenotypic characteristics in close proximity. As expected, T_{FH} cell markers such as CXCR5, PD-1, ICOS, and BCL6 show overlapping expression patterns on the global t-SNE map (**Figure 5.2C**). Notably, while the GC T_{FH} cell subset marker CD57 co-localizes with other T_{FH} cell markers, $CD57^+$ cells segregate into two discrete and well-separated regions (**Figure 5.2D**). The first region (1) is detectable in all three cohorts, whereas the second region (2) shows a distinct pattern of distribution and is most prominently observed in LN cells from HIV^+ samples (**Figure 5.2D**). These data provide the initial evidence to suggest that there may be distinct types of GC T_{FH} cells present at a healthy baseline and under conditions of chronic inflammation.

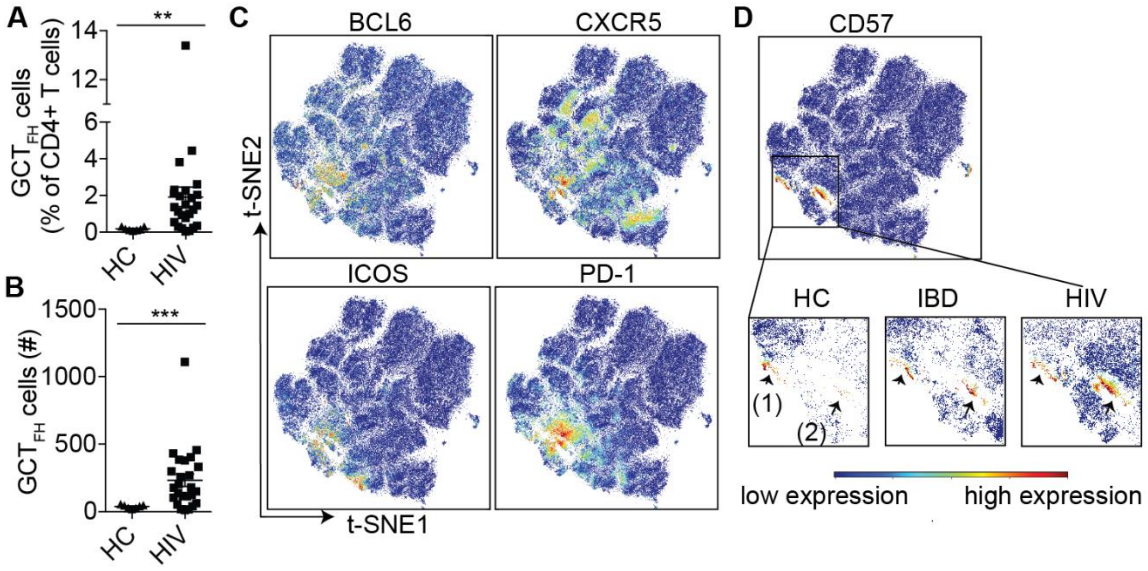

Figure 5.2: High-dimensional analysis of lymphoid CD4$^+$ T cells identified distinct populations of CD4$^+$ T cells with high CD57 expression. (**A**) The frequency of GC T_{FH} cells as a percentage of total CD4$^+$ T cells in the LN. (**B**) The numbers of GC T_{FH} cells detected from 3-5 million CD4$^+$ T cells in each HC or HIV$^+$ sample (HC: n= 7, HIV n = 25). Bar indicates the mean. Statistical significance was analyzed using two-tailed Student's t-test. ** *P < 0.005;* *** *P < 0.0005*. (**C**) CD4$^+$ T cells were analyzed using the viSNE implementation in Cytobank and visualized on a two-dimensional t-SNE map. Data combine samples from all donors and show expression intensity for BCL6, CXCR5, ICOS, and PD-1. (**D**). t-SNE plot showing two discrete regions of high CD57 expression (1) and (2). Inset showing (1) and (2) in equal numbers of CD4$^+$ T cells from HC, IBD, or HIV$^+$ samples. Data and figure produced by L.F.S.

5.2.2 HIV drives expansion of an IL-21-dominant GC T_{FH} phenotype

To further dissect the heterogeneity within GC T_{FH} cells, GC T_{FH} cells were defined according to **Figure D.1** and analyzed by t-SNE (**Figure D.1** and **Figure 5.3A**). The two-dimensional t-SNE map was visualized using a contour plot to facilitate manual gating of discrete cell populations, an approach that has been successfully applied by others to demarcate phenotypically similar clusters[14] (**Figure 5.3B**). This method generated 16 phenotypic clusters. The staining intensity for each phenotypic and

functional marker was visualized as a heatmap for these clusters (**Figure 5.3C**). For comparison, staining intensity is shown for memory cells ($CD45RO^{+}CXCR5^{-}PD1^{-}ICOS^{-}$) and naïve cells ($CD45RO^{-}CXCR5^{-}CD27^{+}CCR7^{+}$) at the bottom of the heatmap. As expected, many GC T_{FH} clusters express classical T_{FH} cell markers, including ICOS, BCL6, and IL-21. However, we also detected a Foxp3 and CD25-expressing cluster that resembles T follicular regulatory cells (T_{FR}) (cluster 13) and a few unexpected clusters expressing granzyme A (clusters 8, 12, 16).

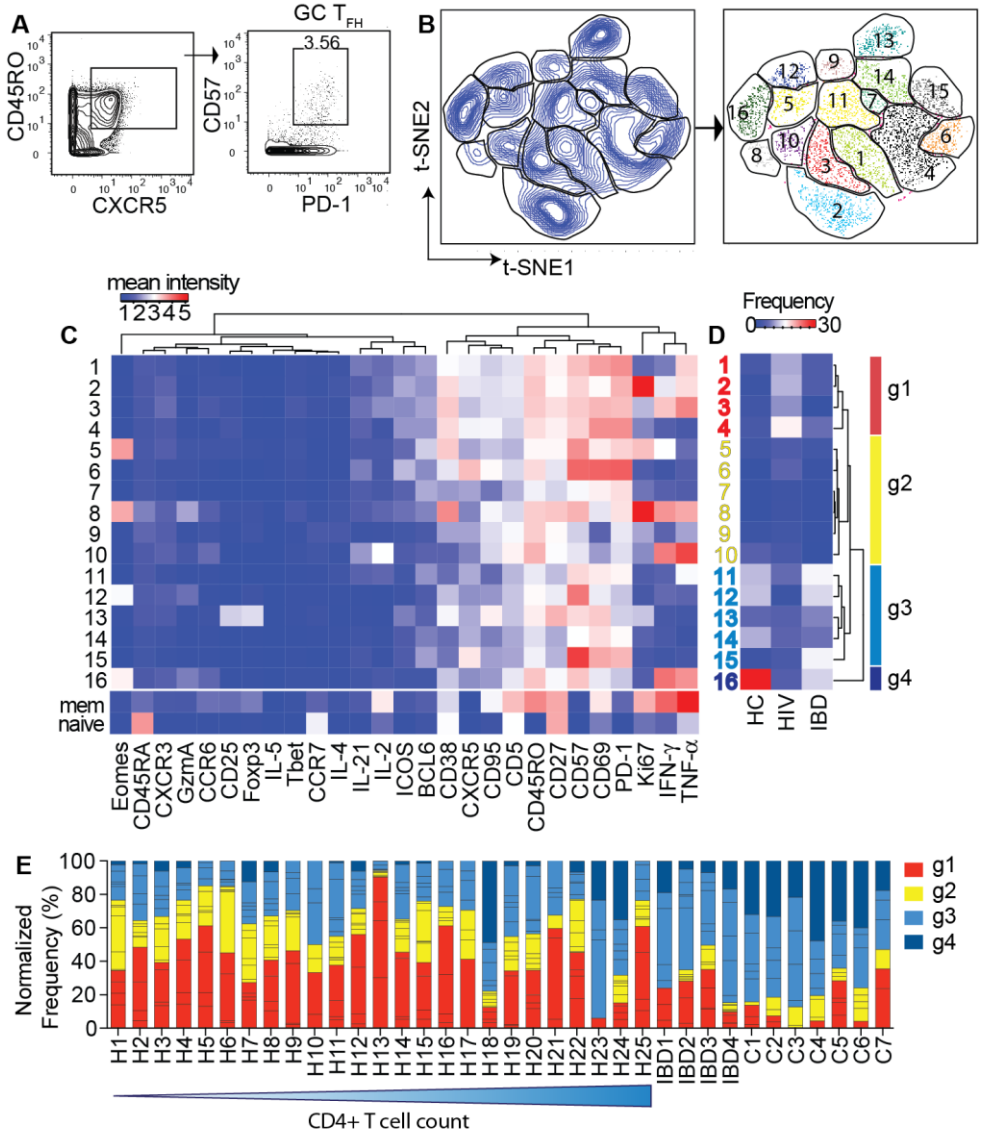
A
CD45RO
CXCR5
GC T FH
3.56
CD57
PD-1
B
t-SNE2
t-SNE1
mean intensity
12345
C
mem
naive
Eomes
CD45RA
CXCR3
GzmA
CCR6
CD25
Foxp3
IL-5
Tbet
CCR7
IL-4
IL-21
IL-2
ICOS
BCL6
CD38
CXCR5
CD95
CD5
CD45RO
CD27
CD57
CD69
PD-1
Ki67
IFN-γ
TNF-α
D
Frequency
0
30
HC
HIV
IBD
g1
g2
g3
g4
E
Normalized Frequency (%)
CD4+ T cell count

Figure 5.3

Figure 5.3: Cellular heterogeneity of GC T_{FH} cells across HC, IBD, and HIV patient-derived LNs. (**A**) Representative gates used to identify GC T_{FH} cells for t-SNE analysis. (**B**) All GC T_{FH} cells from 36 samples were concatenated and displayed on a two-dimensional t-SNE map. Manual gating was performed to identify t-SNE clusters based on contour map (left). Each cluster was assigned an arbitrary number and overlaid onto the t-SNE map (right). (**C**) Heatmap shows raw staining intensity of each marker within each cluster defined in (**B**) after arcsinh transformation. Staining intensities for each marker are shown for memory and naïve cells at the bottom of the heatmap for comparison. (**D**) Heatmap showing the frequency of each cluster in GC T_{FH} cells pooled by disease categories. Cluster groups, g1-g4, are defined by the dendogram (right), which was generated by hierarchical clustering based on cluster frequency in each sample type. (**E**) Stacked bar chart showing normalized frequency of phenotypic group distribution for each sample. HIV^+ samples were ordered by increasing $CD4^+$ T cell count (7 – 1136). Data and figure produced by L.F.S.

To determine which GC T_{FH} cell features represent the baseline state and which are enriched in disease, we pooled GC T_{FH} cells by disease category and measured the percentage of each cluster in HC, IBD, or HIV-patient derived samples. Hierarchical clustering was then applied on cell frequencies with respect to disease state to identify clusters that are shared by individuals in the same clinical category (**Figure 5.3D**). This analysis partitioned the 16 clusters into four major phenotypic groups (g1-g4, **Figure 5.3D**). In general, clusters assigned to groups g1 and g2 were enriched for classical T_{FH} cell markers, such as ICOS and BCL6, and expressed higher levels of IL-21. In contrast, fewer clusters belonging to groups g3 and g4 expressed ICOS, BCL6, or IL-21. Group g3 was comprised mostly of cytokine-low GC T_{FH} cells, including the T_{FR} subset. Group g4 contained cells that predominantly produce non-T_{FH} defining effector molecules, such as TNF-α, IFN-γ, and granzyme A. The distribution of these phenotypic groups was distinct when comparing HC, IBD, and HIV samples. For example, HIV LNs contained the highest frequencies of g1 and lowest frequencies of g3 and g4. IBD LNs were enriched for g3, whereas HC LNs contained the highest frequencies of g4 (**Figure 5.3D**,

Figure D.2). Phenotypic group analysis on individual sample-level data showed a similar pattern with some differences between individuals (**Figure 5.3E**). Notably, our cohort contains three HIV^+ outliers that appeared more HC-like (H18, H23, and H24, **Figure 5.3E**). These samples came from patients who maintained a better $CD4^+$ T cell count without treatment and included an HIV controller with undetectable viral load (H18). Collectively, these data illustrate the heterogeneity of GC T_{FH} cells. While IBD and HIV are both diseases with sustained chronic inflammation, our high dimensional data revealed unique features of GC T_{FH} cells in each disease and highlight the accumulation of a dominant GC T_{FH} population with defined activated features in most HIV^+ patients during chronic HIV infection that are distinct from IBD.

As a result of expanding selected cellular subsets, GC T_{FH} cells in HIV^+ samples may become less functionally diverse. To test this, we examined the breadth of cytokines produced by GC T_{FH} cells. In particular, we focused on the characteristics of IL-21-producing cells to distill the complexity of the dataset, as IL-21 is a key cytokine that mediates T_{FH} cell function. We performed manual gating to identify GC T_{FH} cells that express IL-21 upon stimulation. We then compared $IL\text{-}21^+$ T_{FH} cell frequency in HIV^+ samples to the expected baseline levels from HCs. Consistent with the preferential contribution of HIV^+ samples to the phenotypic group g1, IL-21-expressing GC T_{FH} cells were increased in infected LNs compared to cells from HCs both by percentage and absolute cell number (**Figure 5.4A,B**). To assess poly-functionality, we measured the co-expression of IL-21 with other effector molecules in our antibody panel, which includes IFN-γ, TNF-α, IL-2, IL-4, and granzyme A. Different combinations of cytokine producing subsets were generated by Boolean gating on $IL\text{-}21^+$ GC T_{FH} cells. Cytokine-positive gates were then grouped according to the number of cytokines. This divided GC T_{FH} cells into categories based on the number of effector molecules that were co-

expressed with IL-21, revealing a different distribution of cells in HIV^+ and HC samples. For example, over 40% of IL-21-producing GC T_{FH} cells in HCs simultaneously produced 2 other cytokines (IL-21+2), whereas on average only 18% of cells in the HIV^+ samples were in the same category. Instead, samples from HIV^+ patients were skewed toward single IL-21-producing cells (IL-21 only) (**Figure 5.4C**). To determine if this finding is a general characteristic that extends beyond GC T_{FH} cells, we also examined IL-21-related functional responses in other T_{FH} cell populations. To ensure that we captured additional T_{FH} cells in different stages of differentiation, we broadly defined T_{FH} cells by CXCR5 expression on memory $CD4^+$ T cells. Analysis of IL-21-producing $CXCR5^+$ cells revealed a similar pattern with a statistically significant decrease in IL-21+2 cell subsets and a corresponding increase in IL-21 only cells in HIV^+ samples compared to HCs (**Figure 5.4D**). Notably, poly-functionality of IL-21 producing T_{FH} cells was associated with distinct B cell phenotypes. Past studies have demonstrated fewer memory B cells and higher frequencies of plasma cells in viremic HIV^+ LNs[15, 16]. Here, we identified B cell populations using the same gating strategy as Perreau et al. and found that a low frequency of isotype-switched memory B cells was associated with more single IL-21-producing cells (IL-21 only) and fewer poly-functional (IL-21+2) T_{FH} cells (**Figure D.3**, **Figure 5.4E,F**). A reciprocal relationship was observed for plasma cell frequency, for which a higher frequency was associated with more IL-21 single producers and fewer IL-21+2 T_{FH} cells (**Figure D.4**). Further analysis on individual cytokine-positive populations pointed to a decrease in triple IL-21, IL-2, and TNF-α-producing T_{FH} cells as one of the main differences between HIV^+ and HC LNs (**Figure D.5**). We also compared IL-21-secreting GC T_{FH} cells between HIV and IBD samples to further define similarities and differences under different types of chronic inflammation. This showed a substantial fraction of $IL\text{-}21^-$ single-positive population within IL-21-producing

T_{FH} cells in IBD samples, suggesting that the emergence of a dominant IL-21-producing phenotype may be a shared characteristic of chronic inflammation (**Figure D.6**). However, there were differences between HIV^+ and IBD LNs in other cytokine combinations, including an IL-2 and IL-21-double-positive phenotype that was more frequently expressed in IBD LNs, again highlighting disease-specific features between IBD and HIV infection.

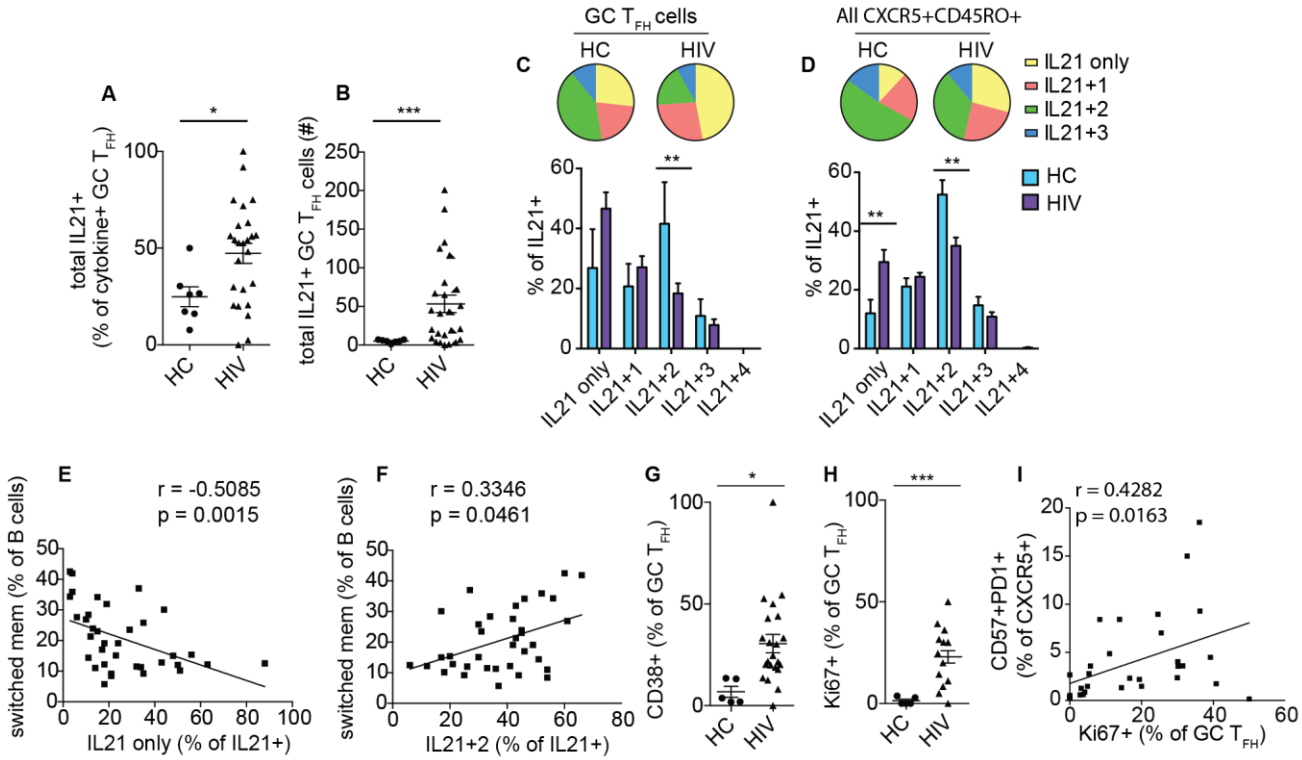

Figure 5.4: IL-21-secreting T_{FH} cells in HIV$^+$ patients acquire an activated phenotype and restricted functional diversity. (**A,B**) The number and frequency of IL21$^+$ T_{FH} cells as a percentage of cytokine producing GC T_{FH} cells. The denominator includes GC T_{FH} cells that produce any combination of IL-2, IFN-γ, TNF-α, IL-4, IL-21, and granzyme A. (**C,D**) Pie charts and bar graphs summarizing the frequency of IL-21$^+$ T_{FH} cells that produce only IL-21 or IL-21 plus 1, 2, 3, or 4 other effector molecules (IL-2, IFN- γ, TNF- α, IL-4, or granzyme A) in the GC subset (**C**) or CXCR5$^+$CD45RO$^+$CD4$^+$ T cells (**D**) from HC or HIV$^+$ LNs. (**E,F**) Correlation between switched memory B cell frequency and the frequency of IL-21 only or IL21+2 producing CXCR5$^+$CD45RO$^+$ cells. Data include all LNs (See also **Figure D.4**). (**G,H**) The frequency of CD38$^+$ or Ki67$^+$ GC T_{FH} cells. (**I**) Correlation between the frequency of CD57$^+$PD1$^+$ cells and Ki67$^+$ GC T_{FH} frequency. Data include all LNs. Statistical analysis using Student's t-test was corrected for multiple comparisons using Holm-Sidak method, with alpha=5%. Error bars represent SEM. Association is measured by Spearman rank correlation and least squares fit regression. * *P < 0.05,* ** *P < 0.005;* *** *P < 0.0005*. Data and figure produced by L.F.S.

5.2.3 GC T_{FH} cells in HIV+ patients have undergone clonal expansion

We next sought to understand the mechanisms that drive the expansion of GC T_{FH} cells during HIV infection. In addition to higher IL-21 level, we also detected higher CD38 and Ki67 expression in GC T_{FH} cells from HIV^+ patients, suggesting that the changes to T_{FH} cells during HIV infection may be induced by T cell activation (**Figure 5.4G,H**). In addition, the frequency of $Ki67^+$ cells correlated with the expression of GC features in the $CXCR5^+$ subset, suggesting that cellular proliferation may contribute to an accumulation of GC T_{FH} cells during chronic HIV infection (**Figure 5.4I**).

How HIV drives GC T_{FH} cell proliferation and expansion remains unclear. The expansion of GC T_{FH} cells may result from an overall non-specific inflammatory state that stimulates a large, diverse group of T_{FH} cells to proliferate. Alternatively, GC T_{FH} proliferation could reflect expansion of a smaller set of HIV-specific T cell clones. To test these different possibilities, we performed TCR sequencing on sorted naïve, $CXCR5^-$ memory, and GC T_{FH} cells directly from LN cell suspension without stimulation. TCRs were sequenced using molecular identifiers (MID) to tag TCR transcripts similar to previously published studies[17, 18] (**Chapter 2**). This effectively increases the accuracy 130 times compared to the first generation of immune repertoire sequencing[19, 20, 21] and also enables accurate quantification of TCR transcript abundance. The number of transcripts detected for a particular CDR3 sequence was then used to define TCR clone size. On average 11,839 TCR transcripts were detected for each sample (**Table D.3**). Unique TCR frequencies range from 1 in 37,129 (0.003%) for the rarest clones to 250 in 2,498 (~10%) for the most expanded clone. To compare the degree of relative clonal expansion, we categorized TCR frequency into 6 groups (rare to >2%) according to the clone size relative to the total TCR transcripts detected in that sample (**Figure 5.5A**). As expected, the TCR repertoire of naïve $CD4^+$ T cells was composed mostly of rare clones

(<0.1%). In contrast, the TCR repertoire of GC T_{FH} cells had a much higher fraction of TCRs occupied by abundant clones (>0.1%) compared to naïve and memory $CD4^+$ T cells (**Figure 5.5A**). The overall diversity of the TCR repertoire was measured by normalized Shannon entropy (NSE), which quantifies the degree of clonal expansion[22, 23]. NSE varies between 0 and 1: 0 if all sequences are the same and 1 if all the sequences are unique. Consistent with the hypothesis that the increase in GC T_{FH} cell frequency is due to selective proliferation of certain T cell clones, NSE analysis revealed that GC T_{FH} cells exhibit significantly more clonal expansion than naive and memory cells (**Figure 5.5B**). Taken together, our data demonstrated a notable expansion of clone size in GC T_{FH} cell populations.

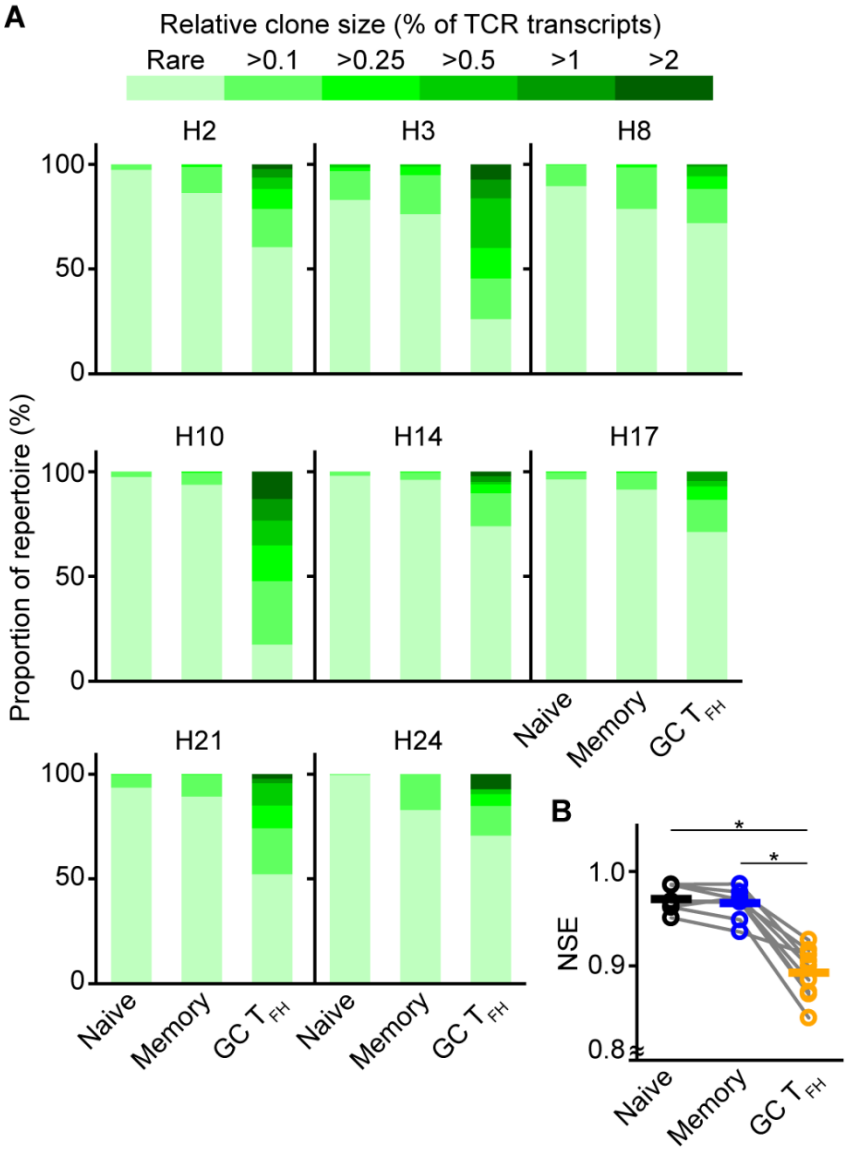

Figure 5.5: GC T_{FH} cells are clonally expanded. (**A**) Breakdown of the proportion of the TCR repertoire represented by clones of different sizes for sorted naïve, memory, and GC T_{FH} cells from HIV^+ LNs. TCR clone size was normalized by the total number of TCR transcripts on nucleotide (nt) sequences, with darker green for TCR clones covering a larger percentage of total number of transcripts and lighter green for TCR clones covering a smaller percentage of total number of transcripts), (**B**) Normalized Shannon entropy of the TCR repertoire of sorted naïve (black), memory (blue), and GC T_{FH} (orange) cells. Grey lines connect samples from the same patient. Sample ID is shown above each bar graph. Bars indicate means. * $P < 0.05$ by two-tailed Wilcoxon signed-rank test.

5.2.4 GC T_{FH} cells show signatures of antigen-driven clonal convergence

The evidence of clonal expansion in GC T_{FH} cells prompted us to ask whether or not the clonal expansion was antigen-driven. To address this question, we analyzed the TCR sequences for evidence of convergence to the same amino acid sequence from distinct nucleotide sequences. Unlike B cells, which can undergo somatic hypermutation, the TCR sequence of a naïve T cell is determined during maturation in the thymus and remains fixed throughout the lifespans of the T cell and its progeny. Thus, with the exception of clones that express 2 TCR α or β sequences, distinct TCR nucleotide sequences necessary arise from distinct naïve T cells. However, multiple nucleotide sequences of different TCRs may encode the same amino acid sequence. These degenerate TCR sequences are typically rare, and the presence of these sequences suggest antigen selection pressure that favors certain TCR motifs that recognize particular antigen(s) (**Figure 5.6B**). Thus, having highly abundant CDR3 amino acid sequences that are encoded by multiple distinct nucleotide sequences indicates preferential expansion of T cells with that specificity[22]. On the other hand, we would not expect multiple nucleotide sequences to converge on the amino acid level in the absence of strong antigen-driven selection. Following this logic, we translated the TCR nucleotide sequences into amino acid sequences and tallied the number of different nucleotide sequences that encode each CDR3 amino acid sequence. These CDR3 amino acid sequences can be broken into 4 quadrants based on the level of degeneracy and frequency in the repertoire (**Figure 5.6A** and **Figure D.7**). Q1 contains highly expanded amino acid CDR3 sequences that are encoded by 2 or more nucleotide sequences. These degenerate, abundant clones likely arose from strong antigen-driven selection and proliferation. Q2 contains low frequency amino acid CDR3 sequences that are also encoded by 2 or more nucleotide sequences. Degenerate clones may originate from naïve T cells, but these are

typically rare as reflected by the low frequency of non-clonally expanded sequences in Q2. Q3 contains amino acid CDR3 sequences that show neither clonal expansion nor amino acid convergence and make up the majority of the repertoire. Q4 contains expanded amino acid CDR3 sequences derived from a single nucleotide sequence and are therefore non-degenerate. This TCR degeneracy analysis revealed a significant degree of antigen-driven clonal convergence in GC T_{FH} cells compared to naïve and memory T cells (**Figure 5.6C**). Together with the NSE decrease in GC T_{FH}, these data provide further evidence that the observed increase in GC T_{FH} frequency was due to antigen-driven clonal expansion.

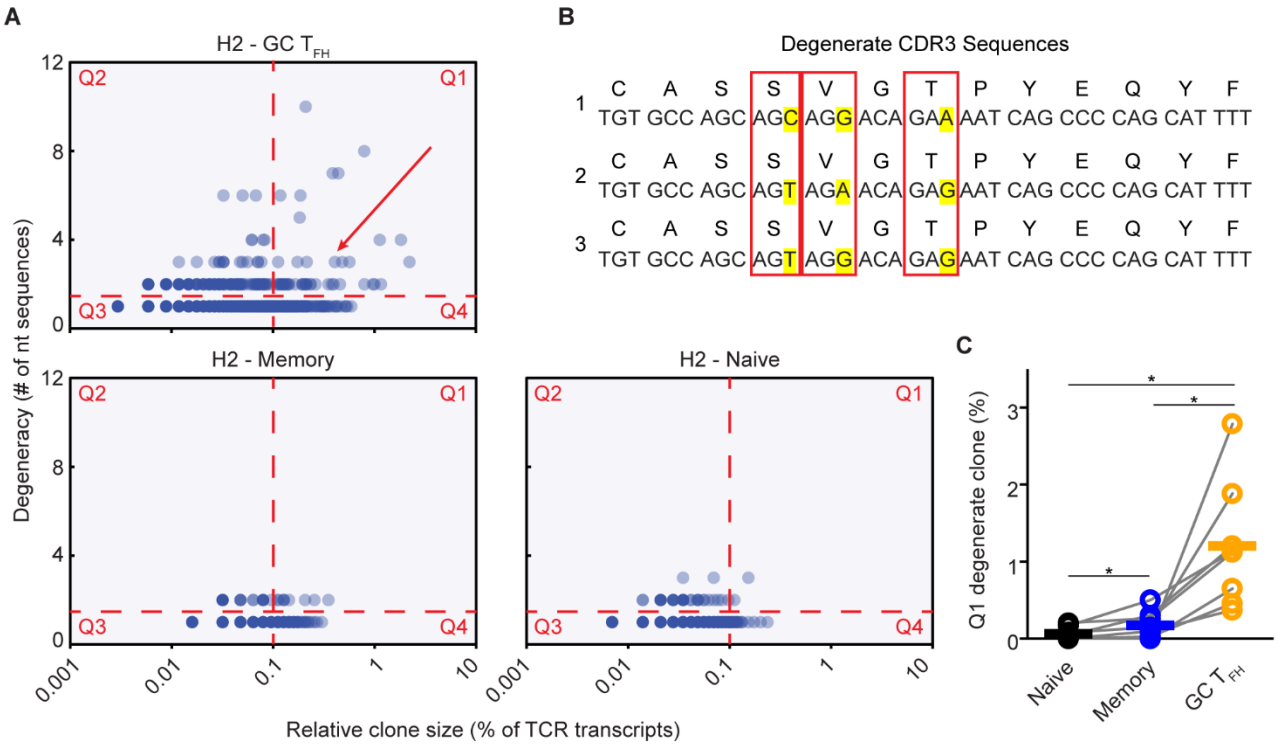

Figure 5.6: Antigen-driven clonal selection signature in GC T_{FH} cells of HIV^+ LNs. (**A**) Representative degeneracy plot from sample H2. Coding degeneracy level (number of unique TCR nucleotide (nt) sequences encoding a common CDR3 amino acid (aa) sequence) of each CDR3 aa sequence is plotted against their frequency (measured as % of total TCR transcript) in naïve, memory, and GC T_{FH} cells. Each dot is a unique CDR3 aa sequence. Red dashed lines indicate cutoffs for degenerate (2 or more nt sequences coding for the same aa sequence, horizontal) and expanded (0.1% or more of TCR transcripts, vertical) clones. Each panel is broken into 4 quadrants: Q1: degenerate-abundant clones; Q2: degenerate-rare clones; Q3: nondegenerate-rare clones; Q4: nondegenerate-abundant clones. Red arrow points to example degenerate clone in (**B**). (**B**) An example of CDR3 aa degeneracy. aa (top row) and nt (bottom row) sequences for each of 3 distinct nt sequences (0.41% of total TCR transcripts) that code for the same aa sequence as the arrow points in (**A**) with Y=3, X=0.41%. Red boxes and highlights indicate redundant codons. (**C**) Comparison of Q1 degenerate-abundant clone percentage in naïve (black), memory (blue), and GC T_{FH} - (orange) cells. Grey lines connect samples from the same patient. Bars indicate means. $*P < 0.05$ by two-tailed Wilcoxon signed-rank test.

5.2.5 GC T_{FH} cells contain HIV-specific T cells

We hypothesized that the antigen-driven clonal expansion observed in GC T_{FH} cells of HIV^+ LNs was caused by HIV viral antigens. To this end, we examined the antigen-specificity of GC T_{FH} cells by establishing donor-specific reference panels of HIV-reactive TCR sequences (**Figure 5.7**). This was performed by culturing LN cells from five HIV^+ patients with an overlapping pool of HIV-1 Gag peptides for 3-4 weeks, then identifying specific antigen-reactive T cells by CD40L and CD69 expression upon peptide restimulation. As a control for non-HIV related clonal expansion, we also cultured and stimulated cells from the same subjects plus one more with overlapping peptides from hemagglutinin (HA) of influenza virus. TCR sequences obtained from these cells were used to identify Gag-specific (**Figure 5.7**) or HA-specific (**Figure D.8**) TCRs from bulk T cell sequences, which were obtained from freshly thawed LN cells that had not been stimulated. By comparing the CDR3 nucleotide sequences from Gag-specific LN cells to bulk naïve, memory, and GC T_{FH} cells, we could trace the overlap in TCR sequences between Gag-specific LN cells and other cell populations. This analysis was illustrated in circos plots (**Figure 5.7A**): the perimeter was divided into four sections corresponding to the four cell populations (**Figure 5.7A**, outer circle). Each thin slice of the arc represents a unique TCR sequence, ordered by the size of that sequence, from the most abundant clones (dark green, 5 or more TCR transcripts) to the rarest clones (light green, single TCR transcript) (**Figure 5.7A**, inner circle). Identical TCR sequences that are shared between the Gag-specific population and naïve, memory, or GC T_{FH} cells were connected by a gray curve.

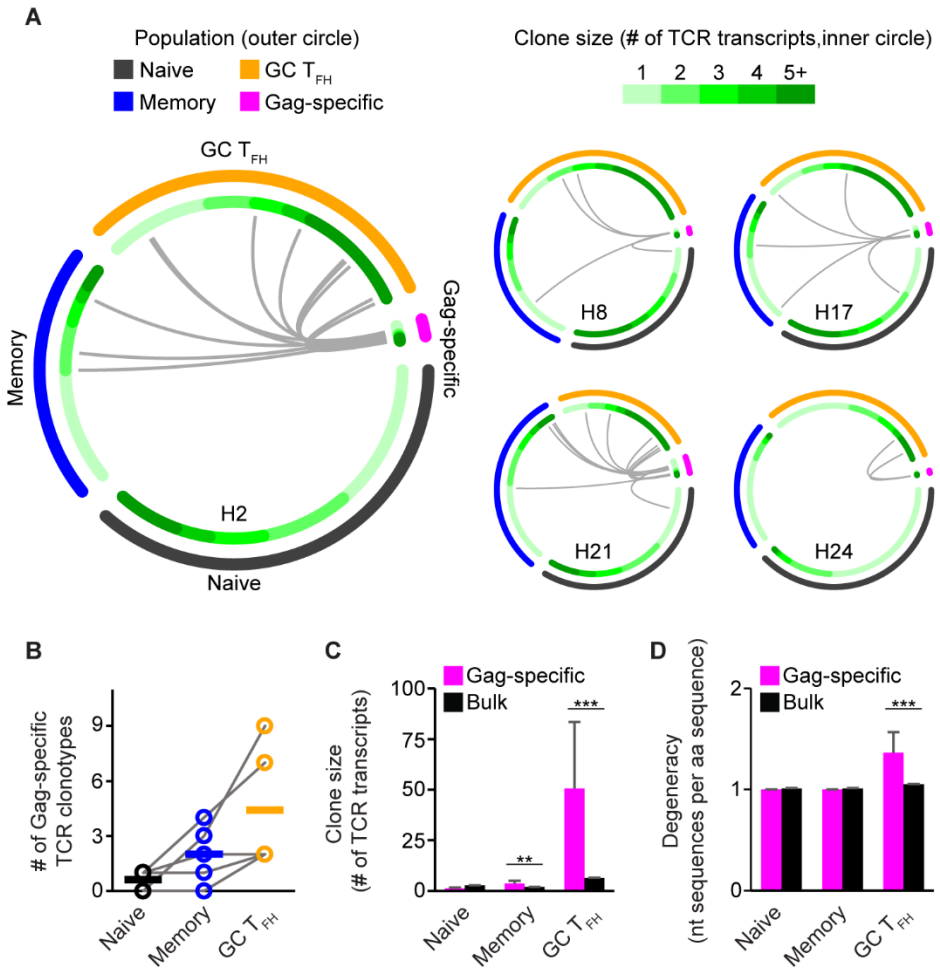

Figure 5.7: GC T_{FH} cells exhibit HIV-antigen-driven clonal expansion and selection. (**A**) Gag-specific TCR clones overlap with HIV^+ LN $CD4^+$ T cell populations. Each thin slice of the arc represents a unique TCR sequence, ordered by the clone size (darker green for larger clones, inner circle). Grey curves indicate Gag-specific TCR clones (nt sequences) found in naïve (black, outer circle), memory (blue, outer circle), and GC T_{FH} (orange, outer circle) populations. (**B**) Number of Gag-specific TCR clones observed in naïve (black), memory (blue), and GC T_{FH} (orange) populations. Grey lines connect samples from the same patient. Bars indicate means. (**C**) Mean clone size of Gag-specific clones (magenta) and clones of unknown specificity (black) from naïve, memory, and GC T_{FH} populations. (**D**) Degeneracy, or the number of distinct nt sequences per CDR3 aa sequence, of Gag-specific clones (magenta) and clones of unknown specificity (black) from naïve, memory, and GC T_{FH} populations. Data from all 5 subjects were aggregated for **C** and **D**. Error bars indicate SEM. ** $P < 0.01$, *** $P < 0.001$ by two-tailed Student's t-test.

We found several Gag-specific TCR sequences in the memory (0 to 4 clones) and GC T_{FH} (2 to 9 clones) populations, while the overlapping was minimum in naïve T cells (0 or 1 clones) (**Figure 5.7B**). Analysis of combined TCR sequencing data from all individuals clearly showed that these Gag-specific GC T_{FH} cells, and to a lesser extent the Gag-specific memory cells, were highly expanded compared to the rest of the T cells with the same phenotypes but unknown specificity (**Figure 5.7C**). As expected, naïve T cells minimally overlap with Gag-specific TCR sequences, and the few sequences that overlapped were not preferentially expanded (**Figure 5.7C**). We also observed sequence overlap between GC T_{FH} cells and HA-reactive T cells (**Figure D.8**). Since influenza exposure contributes to HA-specific memory within circulating T_{FH} cells and most people should have experienced influenza antigen either from natural infection or vaccination[24], it was not surprising that HA-specific T cells overlapped with memory and GC T_{FH} populations. However, in contrast to the observed clonal expansion of Gag-specific T cells within both memory and GC T_{FH} cell compartments, the average clone size for HA-specific T cells is not statistically different from the background level (**Figure D.8C**). Translating these Gag-specific sequences into amino acid sequences and analyzing the occurrence of multiple, distinct nucleotide sequences mapping to the same amino acid sequence showed that the Gag-specific TCR sequences within the GC T_{FH} population, but not naïve nor memory, have a significantly higher degree of coding degeneracy, congruent with our hypothesis that the GC T_{FH} expansion in HIV^+ LNs is antigen-driven (**Figure 5.7D**). These data support a selective expansion of Gag-specific T cells in HIV^+ samples, and collectively provide strong evidence for HIV-antigen-driven changes in GC T_{FH} cells during chronic HIV infection.

5.3 DISCUSSION

Secondary lymphoid organs are the primary sites for immune response against HIV, but they are also preferentially targeted by HIV for viral replication and persistence[25]. Typically, LNs are challenging to access in humans and thus studies using blood and lymphoid tissues from non-human primates have laid the foundation for understanding the immunologic changes during chronic HIV/SIV infection. The data described here represent a comprehensive phenotypic and TCR analysis of GC T_{FH} cells in the LNs from HIV^+ patients and compares the phenotypic changes during HIV infection to other diseased or healthy states using IBD and HC samples. The data show that GC T_{FH} cells are functionally diverse in the LNs of healthy individuals and identify a disruption to the baseline heterogeneity within T_{FH} cells during chronic inflammation. We used TCR repertoire sequencing analysis to further demonstrate virus-driven expansion of HIV-specific GC T_{FH} cells as a key mechanism that contributes to T_{FH} cell pathology in HIV-infected LNs.

One important open question is why an excess of T_{FH} cells in the LNs fails to correct for and may even contribute to impaired humoral immunity during chronic HIV infection. Previous studies have demonstrated functional inhibition of T_{FH} cells by PD-L1 expression on B cells[12]. It has also become clear that there is substantial heterogeneity within T_{FH} cells. For example, Wong et al. used CyTOF to identify many subsets of tonsilar T_{FH} cells from children undergoing tonsillectomy[26]. The biological relevance for T_{FH} cell diversity is just now beginning to be understood. Using IL-21 and IL-4 reporter genes in mice to trace distinct functional subsets T_{FH} cells, Weinstein et al. demonstrated temporal differences in the kinetics of IL-21 and IL-4 production and showed that $IL\text{-}21^+$, $IL\text{-}21^+IL\text{-}4^+$, and $IL\text{-}4^+$ T_{FH} cells each provide specialized follicular helper function and support distinct aspects of B cell function and development[27]. Given these new insights

into the importance of T_{FH} cell subsets, we sought to investigate how cellular heterogeneity of GC T_{FH} cells respond to HIV infection. Past analyses on HIV^+ LNs have generally used a small set of T_{FH} cell markers that appear to be ubiquitously expressed across GC T_{FH} cells, due to the technical limitation of fluorescence-based flow cytometry[12]. Here, we applied mass cytometry to expand the analysis on GC T_{FH} cells using metal isotope-conjugated antibodies against an extended list of surface and intracellular proteins. Our data on LN cells from healthy individuals, IBD patients, and HIV^+ patients revealed a high level of cellular heterogeneity in the GC T_{FH} cell compartment that includes cells with the typical T_{FH} phenotype as well as others that produce non-T_{FH} restricted effector molecules in different combinations.

During chronic HIV infection, GC T_{FH} cells lost functional diversity and acquire a dominant IL-21-producing and activated phenotype. These changes in the functional potential of T_{FH} cells were associated with HIV-related changes in B cell subsets, suggesting that a full complement of T_{FH} cell subsets may be necessary for normal B cell differentiation. Some T_{FH} cell features in infected LNs were shared with LNs from IBD patients who also experience chronic inflammation and impaired humoral response to vaccination[28]. Additionally, there were unique HIV-related changes such as low frequencies of phenotypic groups g3 and g4 that expressed low levels of IL-21. Differences in susceptibility to viral infection and/or virus-induced cytopathology in infected LNs may change the distribution of T_{FH} cell populations. T_{FH} subsets could also compete with and/or regulate each other, either directly or indirectly, by impacting other cell types. Along this line, a major population decreased in HIV^+ LNs produces IL-2, which is a cytokine that antagonizes T_{FH} cell differentiation through Stat5-dependent activation of Blimp-1[29, 30]. Whether IL-2-dependent pathways have a role in maintaining T_{FH} cell diversity is a key question for future studies. Taken together, high-dimensional

CyTOF analysis revealed the diversity of GC T_{FH} cells in LNs. Our data also support the general notion that the cellular environment in HIV^+ LNs is highly pro-inflammatory and additionally highlight the changes in GC T_{FH} cells that are unique to HIV infection.

How HIV promotes changes in GC T_{FH} cell compartment is incompletely understood. The observation of high levels of CD38 and Ki67 expression suggested that GC T_{FH} cells accumulate from T cell activation. However, whether the expansion of GC T_{FH} cells is a result of an overall hyperactive inflammatory state and/or a specific HIV antigen-driven response is not known. Past studies in mice have demonstrated a critical role for sustained antigen availability in the maintenance and differentiation of GC T_{FH} cells[31]. While in these studies TCR engagement is assumed, a direct role for TCRs in GC T_{FH} cell proliferation has not yet been demonstrated. To answer this question, we performed TCR sequencing to precisely define the repertoire composition of GC T_{FH} cells in HIV^+ LNs. Our data revealed that a small portion of these clones harbor distinct nucleotide sequences that converge to the same amino acid sequence – a signature of antigen-driven selection. In addition, convergent TCRs are significantly more common in GC T_{FH} cells compared to memory or naïve cells. These data strongly suggest that HIV-derived antigens directly activate T cells to promote T_{FH} cell pathology in the GC compartment. To test this, we identified HIV-specific TCR sequences from Gag-stimulated T cells from the same LNs. Mapping these HIV-specific TCR sequences back onto the GC T_{FH} TCR repertoire revealed that the Gag-specific T cells in the GC compartment were greatly expanded. The larger clone size and the prevalence of degenerate nucleotide sequences mapping to the same amino acid sequences in HIV-reactive T cells compared to HIV-nonreactive T cells provided further evidence for an HIV-antigen-driven process that leads to the expansion of GC T_{FH} cells during chronic HIV infection.

In conclusion, we identified the expansion of a dominant GC T_{FH} cell population that expresses a defined activated phenotype that is driven by viral antigens during chronic HIV infection. Our combination of high dimensional protein expression and TCR repertoire sequencing analyses provided a powerful platform to interrogate GC T_{FH} cell biology and revealed a previously unappreciated level of heterogeneity and poly-functionality within GC T_{FH} cells at a healthy baseline. Despite their crucial role in GC formation and B cell help, the significantly higher frequencies of GC T_{FH} cells in HIV^+ LNs fail to induce a robust antibody response against the infection. Our data suggest that HIV-driven activation of GC T_{FH} cells in an antigen-specific manner diminishes their poly-functionality, and the emergence of an activated IL-21-constricted phenotype likely contributes to T_{FH} cell dysfunction in severe HIV infection and other conditions of chronic inflammation. Our results highlight the complexity of T_{FH} cells in the LNs and suggest that therapies aiming to restore T_{FH} cell diversity may improve B cell responses in HIV infection.

5.4 METHODS

5.4.1 Human lymph node collection and isolation

The HIV^+ cohort was composed of 25 Mexican individuals. LN excision was obtained from palpable cervical LNs for clinical diagnostic workup and after written informed consent was obtained. Recruitment of study subjects was approved by the Comité de Ciencia y de Ética en Investigación and the Comité de Investigación of the Instituto Nacional de Enfermedades Respiratorias (INER) in México City. LN biopsies from HIV^+ patients were placed in Hanks medium (Lonza) and immediately processed by cutting the tissue in small pieces with a scalpel and cells were dissociated using Gentle MACS tissue dissociator (Miltenyi Biotec). Single cell suspensions were washed and

cryopreserved. All IBD LNs and 2 HC LNs were obtained through Cooperative Human Tissue Network (CHTN) from individuals undergoing clinically indicated bowel resection for IBD or for benign polypectomy. LNs were dissected from peri-colonic adipose tissues shipped overnight in RPMI media. The remaining 4 HC LNs were from iliac region of transplant donors and 1 cervical LN sample was obtained by combining cells from 5 autopsy donors after Ficoll density centrifugation, which was necessary due to poor viability. All samples were de-identified and obtained with IRB regulatory approval from University of Pennsylvania. Subject characteristics are shown in **Table D.2**.

5.4.2 Stimulation and antibody staining for CyTOF

Cryopreserved cells were thawed in 0.025 unit/mL of benzonase (Sigma), washed, and stimulated with 5ng/ml of PMA (Sigma) and 50ng/ml ionomycin (Sigma) in the presence of 2mM Monensin (Sigma) and 5ug/ml Brefeldin A (Sigma) at 37^0C for 5 hours. After stimulation, cells were washed and incubated in 1uM cisplatin (Fluidigm) for 5 min, followed by staining with surface antibody cocktail for 30 min at room temperature (**Table D.1**). Metal conjugation of CyTOF antibodies was performed according to the manufacturer protocol using the X8 Maxpar kit (Fluidigm). For intracellular staining, cells were permeabilized and fixed using Foxp3 staining buffer set (eBioscience) and incubated with the intracellular antibody cocktail for 1 hr at room temperature (**Table D.1**). After staining, cells were washed three times then resuspended in 2% paraformaldehyde (Electron Microscopy Sciences) with 125nM iridium intercalator (Fluidigm) for an overnight incubation at 4^0C. The next day, cells were washed three times, including a final wash in distilled water, and resuspended in PBS containing normalization beads before acquisition on CyTOF 2.

5.4.3 Data analysis for CyTOF

CyTOF FCS files were normalized using bead standards with the Matlab-based Nolan lab normalizer. Iridium^{+}CD3^{+}CD19^{-}TCRαβ^{+}CD4^{+} T cells were excluded for doublets and beads and downsampled to a maximum of 13,910 cells per donor sample using Flowjo's downsampling plug-in to ensure that each sample type contributes equal numbers of CD4^{+} T cells to the data analysis. FCS files were imported into Cytobank for t-SNE analysis using viSNE implementation. For **Figure 5.2**, cluster markers include the following: CD57, CD5, IFN-γ, CCR6, CD69, granzyme A, TNF-α, CD45RO, CD27, Ki67, BCL6, IL-4, CD25, Eomes, ICOS, Foxp3, CD96, CD45RA, CCR7, CXCR5, T-bet, PD-1, IL-21, IL-2, CXCR3, and CD38. ViSNE were set with iteration = 5000, perplexity = 45, theta = 0.5. For the analysis on GC T_{FH} cells, GC T_{FH} cells were defined according to **Figure D.1**. All GC T_{FH} cells were exported from Flowjo and imported into Cytobank using the following implementation (cluster markers: CD57, IFN-γ, CCR6, CD69, GranzymeA, TNF-α, CD27, Ki67, BCL6, CD25, Eomes, ICOS, Foxp3, CD95, CXCR5, PD-1, IL-21, IL-2, CXCR3, CD38, iteration = 2000, perplexity = 30, theta = 0.5). Heatmaps were generated using the "gplot" package in R and show the raw staining intensity of each marker after arcsinh transformation. Heatmap dendograms were clustered by Euclidean distance. CCR5 (156) and BLyS (158) were not included due to significant spill over from isotope impurity by Ki67 (157).

5.4.4 TCRβ library generation and sequencing

Cryopreserved cells were thawed in 0.025 unit/mL of benzonase (Sigma), washed, and stained with CD4-AF700 (OKT4, eBioscience), CD11b-PE/cy5 (ICRF44, Biolegend), CD19-PE/cy5 (HIB19, Biolegend), and CD8-PE/cy5 (HIT8a, Biolegend) in the dump channel, fixable aqua dye (ThermoFisher) for live/dead discrimination, CD57-FITC (HCD57, Biolegend), CD45RO-BV605 (UCHL1, Biolegend), CXC5-PE/TexasRed

(J252D4, Biolegend), ICOS-APC (C398.4, Biolegend), PD-1-BV785 (EH12.2H7, Biolegend), CCR7-PE/cy7 (G043H7, Biolegend). For direct sequencing, thawed cells were sorted by naïve ($CD45RO^-CXCR5^-CD27^+CCR7^+$), memory ($CD45RO^+CXCR5^-PD1^-ICOS^-$), or GC T_{FH} cell phenotypes ($CD45RO^+CXCR5^+PD1^+CD57^+$). To identify antigen-specific T cells, 2-3 million cells were stimulated with 0.5 ug/ml of pooled HA hemagglutinin (HA) peptides from influenza virus (A/California/7/2009) or HIV-1 consensus B gag peptide (Gag 8117, NIH AIDS Reagent Program) in 48-well plates. Cells were cultured for 3-4 weeks, then restimulated with 5 ug/ml of the same peptide pool in the presence of 2 ug/ml of anti-CD40 (HB14, Biolegend). Activated T cells were sorted by positive staining for CD40L (24-31, Biolegend) and CD69 (FN50, BD biosciences). TCR library generation and sequencing on FACS-sorted naïve, memory, GC T_{FH}, and Gag- or HA-reactive T cells was performed using molecular barcodes similar to what has been described previously [17, 18]. Briefly, FACS-sorted cells were lysed, and total RNA was purified using AllPrep DNA/RNA Micro Kit (Qiagen catalog # 80284). 30% of the purified RNA was used for library generation. Most of the libraries were sequenced on Miseq 2x250 PE mode with some sequenced on Hi-seq 2x150 PE mode depending on run availability at the sequencing core. Primers are listed in **Table D.4**.

5.4.5 Sequencing data processing and analysis

Raw reads processing was performed similar to the method previously described in **Chapter 2 Methods 2.4.2-4**. Briefly, consensus sequences were constructed within each MID group. Consensus TCR sequences were subjected to the CDR3 blast module of MIGEC [18] to assign V and J alleles and parse out the CDR3 sequence. Each set of V allele, J allele, and CDR3 sequence is equivalent to a TCR transcript. TCR transcripts

with the same V allele, J allele, and CDR3 sequence were merged to unique TCR sequences, or TCR clones.

5.4.6 Clone size distribution and normalized Shannon entropy

The size of each TCR clone was determined by the number of TCR transcripts of that sequence detected. The sizes were then normalized by the total number of TCR transcripts detected in that sample to yield the relative clone size in percent. These sizes were binned into 6 groups, and the total percent of TCR transcripts represented by each group was displayed in **Figure 5.5A**. Normalized Shannon entropy was calculated as previously described[22, 23].

5.4.7 Amino acid translation and degeneracy

The CDR3 blast module of MIGEC[18] was used to translate the CDR3 nucleotide sequences into amino acid sequences. For each amino acid CDR3 sequence, the number of distinct nucleotide sequences (TCR clones) encoding that amino acid CDR3 sequence was tallied as the degree of degeneracy[22]. Amino acid CDR3 sequences encoded by 2 or more TCR clones were labeled as degenerate, and degeneracy versus relative clone size was analyzed to identify expanded, degenerate clones.

5.4.8 Antigen-specific TCR identification

TCR clones from the peptide-stimulated cells were used to establish donor-specific Gag- and HA-specific TCR sequences. TCR clones found in both the Gag- and HA-stimulated cultures were eliminated, as they likely originated from basally activated T cells within the initial LN sample. These antigen-specific sequences were then queried in the bulk naïve, memory, and GC T_{FH} cells to identify Gag- and HA-specific clones within the respective populations. Circlize R package[32] was used to visualize circos plots in **Figure 5.7A** and **Figure D.8A**. The Gag- and HA-specific clones were isolated from

the rest of the TCR clones, and the clone size and degeneracy were then compared between the antigen-specific TCR clones and the bulk clones (**Figure 5.7C,D** and **Figure D.8C,D**).

5.4.9 Statistics

For pair-wised analysis comparing a single variable in the CyTOF-based phenotypic analysis, statistical significance was analyzed using Student's t-test. Spearman rank correlation and least squares fit regression were applied to measure the degree of association. For TCR repertoire analysis, Wilcoxon signed-rank test was used to compare paired samples, i.e. naïve, memory, and GC T_{FH} cells from matched subjects. A *P*-value of < 0.05 was used as a cutoff for statistical significance. For three-way comparison, significant *P*-value is decreased to < 0.0167. Statistical analyses were performed using GraphPad Prism.

5.5 REFERENCES

1. Vinuesa CG, Linterman MA, Yu D, MacLennan IC. Follicular Helper T Cells. *Annual review of immunology* **34**, 335-368 (2016).

2. Crotty S. T follicular helper cell differentiation, function, and roles in disease. *Immunity* **41**, 529-542 (2014).

3. Kim CH, Rott LS, Clark-Lewis I, Campbell DJ, Wu L, Butcher EC. Subspecialization of CXCR5+ T cells: B helper activity is focused in a germinal center-localized subset of CXCR5+ T cells. *The Journal of experimental medicine* **193**, 1373-1381 (2001).

4. Kim JR, Lim HW, Kang SG, Hillsamer P, Kim CH. Human CD57+ germinal center-T cells are the major helpers for GC-B cells and induce class switch recombination. *BMC immunology* **6**, 3 (2005).

5. Kohler SL*, et al.* Germinal Center T Follicular Helper Cells Are Highly Permissive to HIV-1 and Alter Their Phenotype during Virus Replication. *Journal of immunology* **196**, 2711-2722 (2016).

6. Hufert FT, *et al.* Germinal centre CD4+ T cells are an important site of HIV replication in vivo. *Aids* **11**, 849-857 (1997).

7. Haas A, Zimmermann K, Oxenius A. Antigen-dependent and -independent mechanisms of T and B cell hyperactivation during chronic HIV-1 infection. *Journal of virology* **85**, 12102-12113 (2011).

8. De Milito A, *et al.* Mechanisms of hypergammaglobulinemia and impaired antigen-specific humoral immunity in HIV-1 infection. *Blood* **103**, 2180-2186 (2004).

9. Parmigiani A, *et al.* Impaired antibody response to influenza vaccine in HIV-infected and uninfected aging women is associated with immune activation and inflammation. *PLoS One* **8**, e79816 (2013).

10. Crum-Cianflone NF, *et al.* Durability of antibody responses after receipt of the monovalent 2009 pandemic influenza A (H1N1) vaccine among HIV-infected and HIV-uninfected adults. *Vaccine* **29**, 3183-3191 (2011).

11. Tebas P, *et al.* Poor immunogenicity of the H1N1 2009 vaccine in well controlled HIV-infected individuals. *Aids* **24**, 2187-2192 (2010).

12. Rafael AC, *et al.* Inadequate T follicular cell help impairs B cell immunity during HIV infection. *Nat Med* **19**, 494-499 (2013).

13. Spitzer MH, Nolan GP. Mass Cytometry: Single Cells, Many Features. *Cell* **165**, 780-791 (2016).

14. Wong MT, *et al.* A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures. *Immunity* **45**, 442-456 (2016).

15. Lindqvist M, *et al.* Expansion of HIV-specific T follicular helper cells in chronic HIV infection. *The Journal of clinical investigation* **122**, 3271-3280 (2012).

16. Matthieu P, *et al.* Follicular helper T cells serve as the major CD4 T cell compartment for HIV-1 infection, replication, and production. *The Journal of experimental medicine* **210**, 143-156 (2013).

17. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the*

National Academy of Sciences of the United States of America **110**, 13463-13468 (2013).

18. Shugay M*, et al.* Towards error-free profiling of immune repertoires. *Nature methods*, (2014).

19. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807-810 (2009).

20. Jiang N, Weinstein JA, Penland L, White RA, 3rd, Fisher DS, Quake SR. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5348-5353 (2011).

21. Jiang N*, et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine* **5**, 171ra119 (2013).

22. Jia Q*, et al.* Diversity index of mucosal resident T lymphocyte repertoire predicts clinical prognosis in gastric cancer. *Oncoimmunology* **4**, e1001230 (2015).

23. Lin J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* **37**, 145-151 (1991).

24. Herati RS*, et al.* Successive annual influenza vaccination induces a recurrent oligoclonotypic memory response in circulating T follicular helper cells. *Science Immunology* **2**, (2017).

25. Lorenzo-Redondo R*, et al.* Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature* **530**, 51-56 (2016).

26. Wong MT*, et al.* Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell reports* **11**, 1822-1833 (2015).

27. Weinstein JS*, et al.* TFH cells progressively differentiate to regulate the germinal center response. *Nature immunology* **17**, 1197-1205 (2016).

28. Marin AC, Gisbert JP, Chaparro M. Immunogenicity and mechanisms impairing the response to vaccines in inflammatory bowel disease. *World journal of gastroenterology* **21**, 11273-11281 (2015).

29. Oestreich KJ, Mohn SE, Weinmann AS. Molecular mechanisms that control the expression and activity of Bcl-6 in TH1 cells to regulate flexibility with a TFH-like gene profile. *Nature immunology* **13**, 405-411 (2012).

30. Johnston RJ, Choi YS, Diamond JA, Yang JA, Crotty S. STAT5 is a potent negative regulator of TFH cell differentiation. *The Journal of experimental medicine* **209**, 243-250 (2012).

31. Baumjohann D*, et al.* Persistent antigen and germinal center B cells sustain T follicular helper cell responses and phenotype. *Immunity* **38**, 596-605 (2013).

32. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).

Chapter 6: Conclusions and Future Studies

We developed an improved IR-Seq method, MIDCIRS, and applied it to study the antibody repertoire response to malaria in young children and the perturbation of the T_{FH} compartment during HIV infection. MIDCIRS utilizes a 12N barcode to track individual mRNA transcripts through PCR amplification and sequencing. Due to the stochastic nature of the randomized barcodes, especially for samples containing a large number of transcripts, distinct mRNA transcripts may happen to be labeled with identical barcodes. We performed a sequence-similarity-based clustering step to recognize these occurrences and further separate MID groups into sub-groups, with each sub-group representing an mRNA molecule. Using naive B cells at varying cell counts, we developed a set of metrics than can be used to validate future IR-Seq methods and experiments. We showed that MIDCIRS has excellent diversity coverage down to as few as 1,000 cells, setting it apart from other IR-Seq methods. MIDCIRS is extremely useful for analyzing precious samples involving low blood draw volumes and/or rare cell types.

This aspect of MIDCIRS made it perfect to interrogate the infant and toddler antibody repertoire. The predictable annual malaria season in Western Africa opened up the possibility of studying the age-related development of the antibody repertoire and its capacity to respond to a natural infection. We found that the infant antibody repertoire is surprisingly competent. While the overall SHM load is low in infants, they have the ability to produce a small fraction of highly mutated antibodies. Upon acute malaria infection, infants significantly diversify their B cell clonal lineages, on par with what we observed in toddlers. The high coverage of MIDCIRS allowed for us to discover antibody sequences during acute malaria that could be traced back to memory B cells from the pre-malaria timepoint in toddlers who had previously experienced malaria. The acute progeny

of these memory B cells harbored significantly more mutations, and we found isotype switched sequences stemming from IgM-expressing memory B cells, highlighting the ability of memory B cells to further induce SHM and class switch during a secondary infection.

Despite the apparent competence of the infant antibody repertoire, children in malaria-endemic regions fail to develop sterile immunity to malaria and only develop clinical immunity around the age of 7 years old. It will be of great interest to extend this antibody repertoire analysis to older children to pinpoint differences that lead to clinical immunity. A longitudinal cohort combined with the multiple timepoint lineage approach we showcased could trace the evolution of B cell clonal lineages that contribute to this immunity, possibly shedding light on why repeated infections are necessary to establish clinical immunity to malaria. Understanding the mechanism of natural immunity to malaria could be a key step towards improving malaria vaccine strategies, as the current best vaccine only offers limited protection, especially for infants.

MIDCIRS reduces the error rate low enough to confidently call mutations from sequencing errors, but it cannot account for any discrepancies between the reference germline sequence and an individual's germline sequence. Such mismatches would inflate SHM counts and lead to skewed repertoire analysis. We developed a simple yet effective technique for bioinformatically predicting novel germline alleles from antibody repertoire sequencing data and validating them experimentally. Each of the 8 individuals tested had at least 1 novel germline allele sequence, demonstrating the need to perform this type of correction. This approach should be used for any and all antibody repertoire sequencing analysis to instill confidence in SHM calling.

We applied MIDCIRS to study the T_{FH} TCR repertoire in HIV patients. Despite decimating the circulating $CD4^+$ T cell population, HIV infection causes expansion of

functionally restricted T_{FH} cells within secondary lymphoid tissues. These functionally restricted T_{FH} cells fail to induce a neutralizing antibody response to HIV. We showed that these expanded T_{FH} cells were preferentially HIV-specific and showed signs of antigen-driven convergent evolution. It remains to be seen if interventions that restore polyfunctionality to these T_{FH} cells can lead to a more robust antibody response.

In future studies, the exquisite sensitivity and coverage of MIDCIRS could be useful not only for research applications but also for clinical diagnostics. B and T cells play a prominent role in many diseases. The enhanced ability to detect rare clones may prove useful for applications such as early diagnostic of autoimmune diseases. The ability to extract high quality sequencing data from minimal cell inputs may prove useful for TCR repertoire analysis of tumor infiltrating lymphocytes. Overall, IR-Seq is still in the infancy stage, and MIDCIRS adds another tool to the toolbox for researchers to expand our understanding of the immune system.

Appendix A – Chapter 2 Supplementary Information

Primer name	Sequence (5'-3')
RT primers	
IgG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNN NNNAAGACCGATGGGCCCTTG
IgA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNN NNNGAAGACCTTGGGGCTGGT
IgM	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNN NNNGGGAATTCTCACAGGAGACG
IgE	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNN NNNGAAGACGGATGGGCTCTGT
IgD	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNN NNNGGGTGTCTGCACCCTGATA
1st PCR forward primers	
ILLUPE2LR1	GACGTGTGCTCTTCCGATCTCGCAGACCCTCTCACTCAC
ILLUPE2LR2	GACGTGTGCTCTTCCGATCTTGGAGCTGAGGTGAAGAAGC
ILLUPE2LR3	GACGTGTGCTCTTCCGATCTTGCAATCTGGGTCTGAGTTG
ILLUPE2LR4	GACGTGTGCTCTTCCGATCTGGCTCAGGACTGGTGAAGC
ILLUPE2LR5	GACGTGTGCTCTTCCGATCTTGGAGCAGAGGTGAAAAAGC
ILLUPE2LR6	GACGTGTGCTCTTCCGATCTGGTGCAGCTGTTGGAGTCT
ILLUPE2LR7	GACGTGTGCTCTTCCGATCTACTGTTGAAGCCTTCGGAGA
ILLUPE2LR8	GACGTGTGCTCTTCCGATCTAAACCCACACAGACCCTCAC
ILLUPE2LR9	GACGTGTGCTCTTCCGATCTAGTCTGGGGCTGAGGTGAAG
ILLUPE2LR10	GACGTGTGCTCTTCCGATCTGGCCCAGGACTGGTGAAG
ILLUPE2LR11	GACGTGTGCTCTTCCGATCTGGTGCAGCTGGTGGAGTC
1st PCR reverse primer	
ILLUPE1adaptor_short	ACACTCTTTCCCTACACGAC
2nd PCR reverse primer	
ILLUPE1adaptor	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC GAC
2nd PCR forward primers with 7 library barcodes	
ILLUPE2TSBC21	CAAGCAGAAGACGGCATACGAGATAACGAAACGTGACTGGA GTTCAGACGTGTGCTCTTCCGATCT
ILLUPE2TSBC22	CAAGCAGAAGACGGCATACGAGATAACGTACGGTGACTGGA GTTCAGACGTGTGCTCTTCCGATCT
ILLUPE2TSBC23	CAAGCAGAAGACGGCATACGAGATAACCACTCGTGACTGGA GTTCAGACGTGTGCTCTTCCGATCT
ILLUPE2TSBC25	CAAGCAGAAGACGGCATACGAGATAAATCAGTGTGACTGGA

	GTTCAGACGTGTGCTCTTCCGATCT
ILLUPE2TSBC26	CAAGCAGAAGACGGCATACGAGATAAGCTCATGTGACTGGA GTTCAGACGTGTGCTCTTCCGATCT
ILLUPE2TSBC27	CAAGCAGAAGACGGCATACGAGATAAAGGAATGTGACTGGA GTTCAGACGTGTGCTCTTCCGATCT
ILLUPE2TSBC28	CAAGCAGAAGACGGCATACGAGATAACTTTTGGTGACTGGA GTTCAGACGTGTGCTCTTCCGATCT

Table A.1: Primers used for antibody sequence library generation.

Appendix B – Chapter 3 Supplementary Information

Patient	Pre-malaria				Acute malaria		
	Pre-Index	Pre-Age	PBMC	Memory B	Acute-Index	Acute Age	PBMC
Inf1	Inf1-Pre3m	3m	Yes	I.S.	Inf1-Acu9m	9m	Yes
Inf2	Inf2-Pre3m	3m	Yes	J.F.	Inf2-Acu6m	6m	Yes
Inf3	Inf3-Pre5m	5m	Yes	I.S.	Inf3-Acu11m	11m	Yes
Inf4	Inf4-Pre5m	5m	Yes	J.F.	Inf4-Acu10m	10m	Yes
Inf5*	Inf5-Pre5m	5m	Yes	J.F.	Inf5-Acu10m	10m	Yes
Inf6	Inf6-Pre8m	8m	Yes	J.F.	Inf6-Acu12m	12m	Yes
Inf7	Inf7-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf8	Inf8-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf9	Inf9-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf10	Inf10-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Inf11	Inf11-Pre11m	11m	Yes	Yes	N.A.	N.A.	N.A.
Tod1*	Tod1-Pre17m	17m	Yes	Yes	Tod1-Acu22m	22m	Yes
Tod2	Tod2-Pre19m	19m	Yes	Yes	Tod2-Acu22m	22m	Yes
Tod3†	Tod3-Pre28m	28m	Yes	Yes	Tod3-Acu32m	32m	Yes
Tod4	Tod4-Pre29m	29m	Yes	Yes	Tod4-Acu32m	32m	Yes
Tod5	Tod5-Pre31m	31m	Yes	J.F.	Tod5-Acu32m	32m	Yes
Tod6	Tod6-Pre31m	31m	Yes	Yes	Tod6-Acu38m	38m	Yes
Tod7†	Tod7-Pre40m	40m	Yes	Yes	Tod7-Acu42m	42m	Yes
Tod8	Tod8-Pre42m	42m	Yes	Yes	Tod8-Acu46m	46m	Yes
Tod9	Tod9-Pre47m	47m	Yes	Yes	Tod9-Acu50m	50m	Yes
Tod10	Tod10-Pre13m	13m	Yes	Yes	N.A.	N.A.	N.A.
Tod11	Tod11-Pre16m	16m	Yes	Yes	N.A.	N.A.	N.A.
Tod12	Tod12-Pre17m	17m	Yes	Yes	N.A.	N.A.	N.A.
Tod13	Tod13-Pre17m	17m	Yes	Yes	N.A.	N.A.	N.A.

Table B.1: Cohort and cell type availability. I.S. indicates insufficient PBMC for FACS sorting or analysis. J.F. indicates just flow cytometry analysis. N.A indicates samples were not available. * Same individual. † Same individual.

Sample	PBMCs[a]	Raw reads	Mapped reads	Percent Mapped	Unique RNA molecules
Inf1-Pre3m	3,000,000	3,246,180	2,989,252	92.1%	41,842
Inf1-Acu9m	3,000,000	3,608,436	3,348,589	92.8%	32,800
Inf2-Pre3m	3,000,000	3,176,623	2,987,587	94.0%	35,379
Inf2-Acu6m	3,000,000	3,689,115	3,481,675	94.4%	29,523
Inf3-Pre5m	4,150,000	3,242,619	3,070,458	94.7%	37,234
Inf3-Acu11m	5,000,000	4,396,739	4,153,830	94.5%	42,634
Inf4-Pre5m	5,000,000	3,048,762	2,810,018	92.2%	45,445
Inf4-Acu10m	3,700,000	5,287,767	4,864,629	92.0%	29,694
Inf5-Pre5m*	5,000,000	3,764,663	3,425,015	91.0%	54,516
Inf5-Acu10m*	50,00,000	4,712,120	4,374,600	92.8%	41,774
Inf6-Pre8m	5,000,000	3,588,177	3,456,165	96.3%	47,254
Inf6-Acu12m	400,000	395,765	378,182	95.6%	03,447
Tod1-Pre17m*	5,000,000	2,816,309	2,576,372	91.5%	53,551
Tod1-Acu22m*	1,380,000	2,811,617	2,593,849	92.3%	12,514
Tod2-Pre19m	5,000,000	4,842,338	4,673,875	96.5%	40,600
Tod2-Acu22m	1,920,000	1,956,906	1,886,521	96.4%	15,285
Tod3-Pre28m†	5,000,000	3,988,677	3,687,883	92.5%	35,567
Tod3-Acu32m†	5,000,000	9,218,255	8,565,149	92.9%	47,144
Tod4-Pre29m	5,000,000	2,924,629	2,851,964	97.5%	48,950
Tod4-Acu32m	5,000,000	4,004,416	3,846,197	96.0%	40,628
Tod5-Pre31m	5,000,000	5,338,867	5,126,888	96.0%	31,531
Tod5-Acu32m	3,000,000	2,853,984	2,736,902	95.9%	26,955
Tod6-Pre31m	5,000,000	4,356,975	4,198,929	96.4%	44,665
Tod6-Acu38m	2,170,000	5,738,001	5,460,964	95.2%	22,270
Tod7-Pre40m†	5,000,000	3,192,503	2,893,482	90.6%	34,901
Tod7-Acu42m†	4,740,000	4,448,008	4,079,432	91.7%	34,185
Tod8-Pre42m	5,000,000	2,120,127	2,058,164	97.1%	48,939
Tod8-Acu46m	2,100,000	2,060,234	1,986,239	96.4%	17,039
Tod9-Pre47m	3,000,000	3,035,618	2,682,991	88.4%	20,094
Tod9-Acu50m	3,000,000	4,678,879	3,912,981	83.6%	18,447

Table B.2: Sequencing reads statistics of paired PBMCs from the malaria cohort. [a]Number of PBMCs differs because of the age dependent blood draw volume and cell recovery. * Same individual. † Same individual.

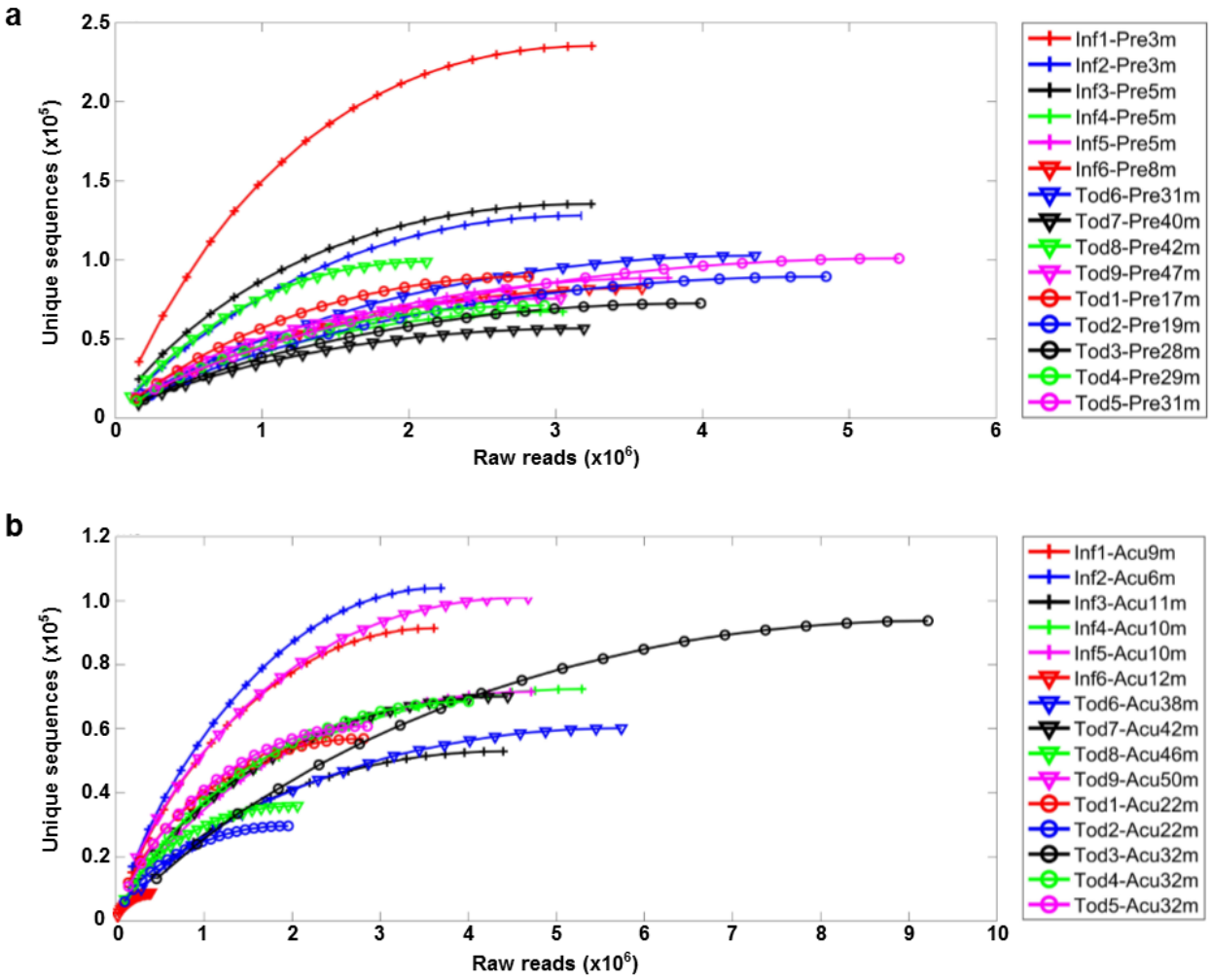

Figure B.1: Rarefaction analysis of paired PBMC malaria cohort sequencing libraries. (**a**) Pre-malaria PBMC rarefaction curves (N=15). (**b**) Acute malaria PBMC rarefaction curves (N=15). Raw reads were subsampled to varying depths, and MIDCIRS was used to determine the number of unique RNA molecules. All single-read sequences that occurred before subsampling were discarded. Single-read sequences that occurred as a results of subsampling were included as unique RNA molecules. The number of unique RNA molecules discovered saturated for all samples, indicating adequate sequencing depth.

Sample	Percentage of Unique RNA sequences assigned to novel germline alleles
Inf1-Pre3m	4.81%
Inf1-Acu9m	6.21%

Inf2-Pre3m	8.44%
Inf2-Acu6m	9.11%
Inf3-Pre5m	1.78%
Inf3-Acu11m	4.91%
Inf4-Pre5m	11.83%
Inf4-Acu10m	9.63%
Inf5-Pre5m*	8.19%
Inf5-Acu10m*	7.72%
Inf6-Pre8m	6.02%
Inf6-Acu12m	6.79%
Tod1-Pre17m*	9.82%
Tod1-Acu22m*	7.51%
Tod2-Pre19m	2.54%
Tod2-Acu22m	2.34%
Tod3-Pre28m†	16.91%
Tod3-Acu32m†	15.05%
Tod4-Pre29m	3.61%
Tod4-Acu32m	4.80%
Tod5-Pre31m	6.98%
Tod5-Acu32m	6.79%
Tod6-Pre31m	5.89%
Tod6-Acu38m	4.15%
Tod7-Pre40m†	18.30%
Tod7-Acu42m†	13.84%
Tod8-Pre42m	7.40%
Tod8-Acu46m	5.71%
Tod9-Pre47m	13.10%
Tod9-Acu50m	13.15%

Table B.3: Percentage of unique RNA sequences assigned to novel alleles for each sample. Novel alleles detected by TIgGER and our method were combined. * Same individual. † Same individual.

Sample	Number of naive B cells	Average number of mutations
Inf1-Acu9m	10000	0.31
Inf2-Pre3m	10000	0.20
Inf3-Pre5m	10000	0.29
Inf4-Pre5m	10000	0.27
Inf5-Pre5m*	10000	0.40
Tod1-Pre17m*	10000	0.79
Tod2-Pre19m	10000	0.57
Tod3-Pre28m†	10000	0.53
Tod4-Pre29m	100000	1.07
Tod7-Pre40m†	10000	0.45
Tod8-Pre42m	100000	1.20

Table B.4: Average mutation number of naive B cells. * Same individual. † Same individual.

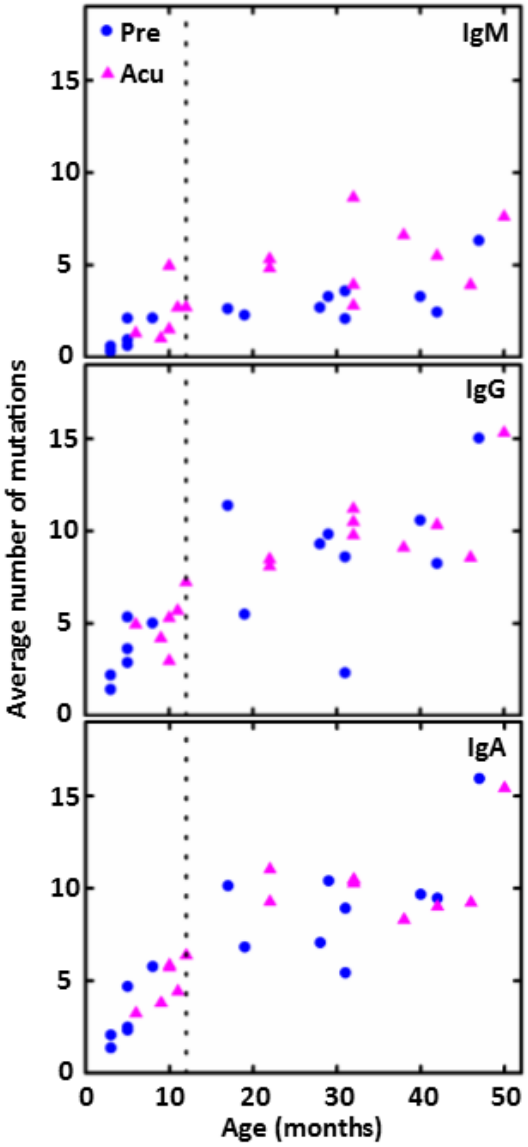

Figure B.2: Correlation between average number of mutations and age for initial, paired pre- and acute malaria samples. Initial samples (N=15) suggested a step-wise increase in SHM load around 12 months which prompted us to divide our cohort into two age groups and delve further into the antibody repertoire properties. We since added 9 pre-malaria samples around the transition, 11 months to 17 months, which were shown in **Figure 3.5**.

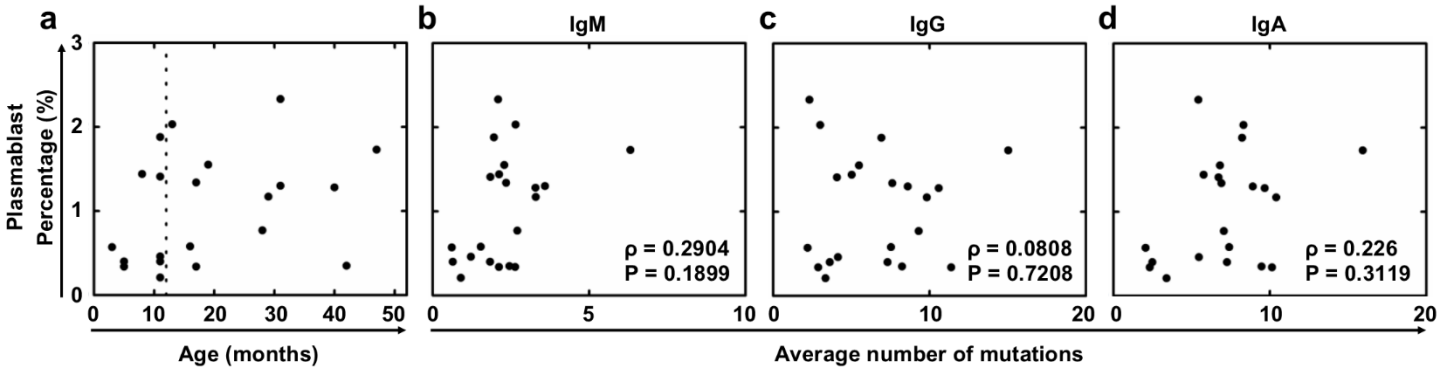

Figure B.3: Comparison between pre-malaria plasmablast percentage of total B cells and average number of mutations. (**a**) Plasmablast percentages of total B cells compared with age. (**b-d**) Plasmablast percentages of total B cells compared with average number of mutations of IgM (**b**), IgG (**c**), and IgA (**d**) sequences from bulk PBMCs in pre-malaria samples from infants (N=9) and toddlers (N=13). ρ and P values determined by Spearman's rank correlation have been listed in the figure.

			FWR			CDR			Average R/S Ratio	
			R	S	R/S Ratio	R	S	R/S Ratio	FWR	CDR
Infant	**Pre**	IgM	0.54	0.11	4.98	0.18	0.04	5.15	3.00 ± 1.12	5.54 ± 0.25
		IgG	1.54	0.70	2.21	1.36	0.24	5.67		
		IgA	1.48	0.65	2.28	1.29	0.22	5.75		
	Acute	IgM	1.36	0.34	4.05	0.58	0.11	5.52		
		IgG	1.88	0.85	2.22	1.62	0.30	5.35		
		IgA	2.03	0.90	2.25	1.75	0.30	5.79		
Toddler	**Pre**	IgM	1.12	0.35	3.20	0.58	0.11	5.54	2.41 ± 0.45	5.34 ± 0.25
		IgG	3.42	1.57	2.17	2.73	0.54	5.05		
		IgA	3.88	1.82	2.14	3.15	0.58	5.41		
	Acute	IgM	2.16	0.79	2.73	1.33	0.24	5.44		
		IgG	4.28	2.02	2.11	3.39	0.68	5.02		
		IgA	4.33	2.04	2.12	3.55	0.64	5.59		

Table B.5: Replacement and silent mutations and their ratios for PBMCs in infants and toddlers. Nucleotide mutations resulting in amino acid substitutions (Replacement, R) or no amino acid substitutions (silent, S) in the framework region (FWR2 and 3) and complementary determining regions (CDR1 and 2) of infants (N=6) and toddlers (N=9), weighted by unique RNA molecules. CDR3 and FWR4 were not included in this analysis due to the difficulty determining the germline sequence. FWR1 for all sequences was also omitted because it was not covered entirely by some of the primers. Average displayed as mean ± standard deviation.

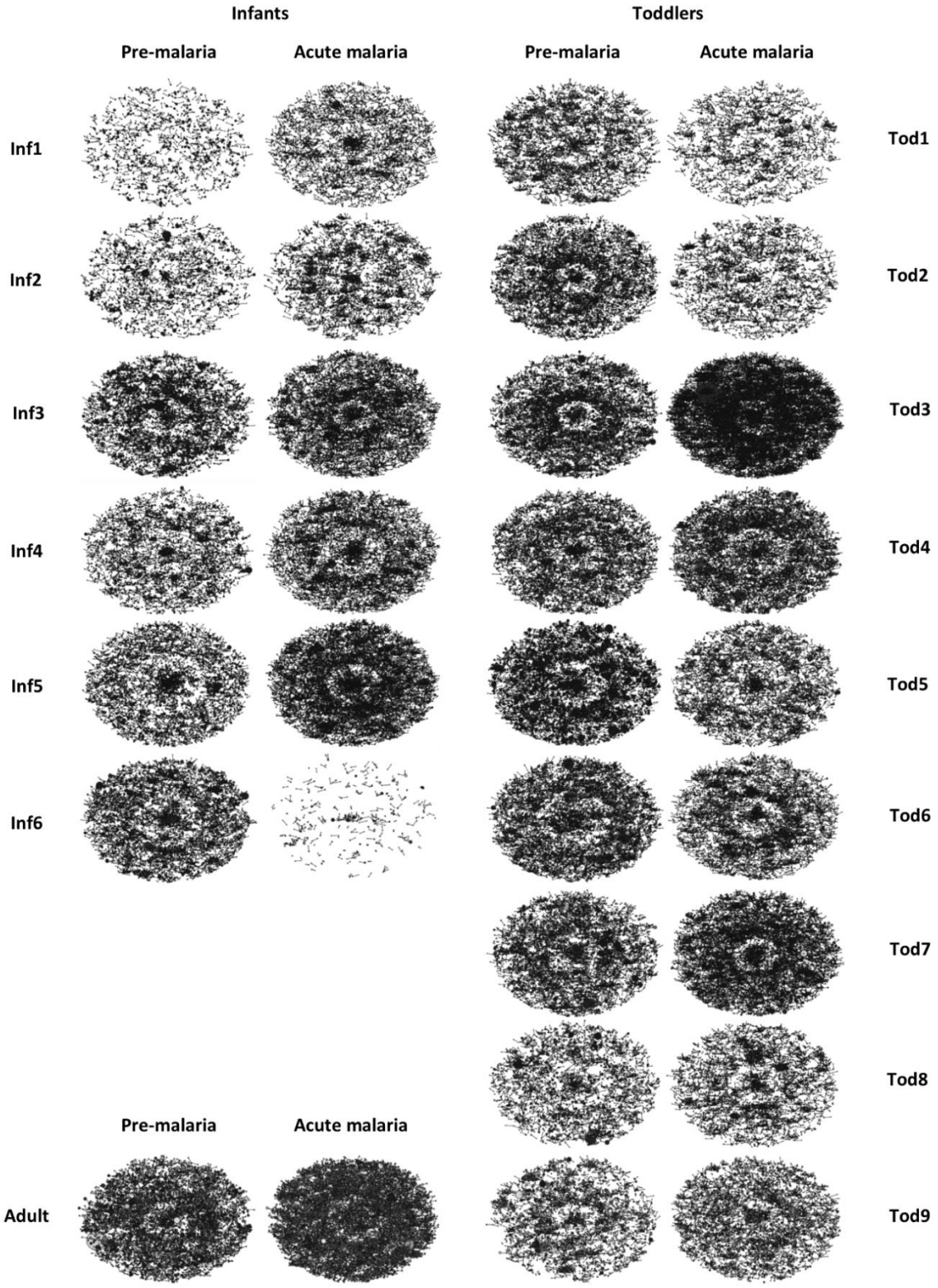
Infants
Toddlers
Pre-malaria
Acute malaria
Pre-malaria
Acute malaria
Inf1
Tod1
Inf2
Tod2
Inf3
Tod3
Inf4
Tod4
Inf5
Tod5
Inf6
Tod6
Tod7
Tod8
Pre-malaria
Acute malaria
Adult
Tod9

Figure B.4

Figure B.4: Lineage structure visualization. Lineage distribution structures for pre-malaria and acute malaria samples for all individuals with corresponding pre-malaria and acute malaria PBMC samples. A 24 year old adult malaria patient was also included. Lineages composed of only a single unique RNA molecule were excluded. Clonal lineages shown in **Figure 3.10** are densely packed here. Therefore, it is not intended to show intra-lineage structure for all individual lineages in each panel; rather, each panel provides an overview of all lineages for one individual at one timepoint. The darker the cluster in each oval-shaped global lineage map, the more densely packed lineages there are.

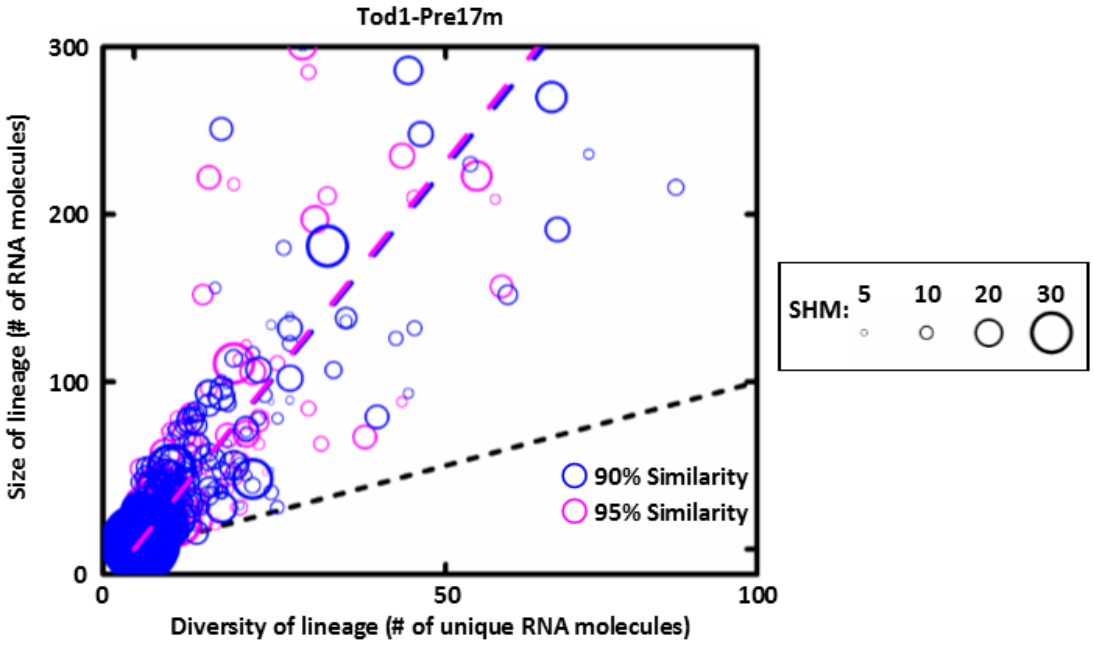

Figure B.5: Comparison between different thresholds for lineage formation. 90% (blue) and 95% (pink) nucleotide similarities of the CDR3 region were used as the threshold to generate lineages. The distribution of the size vs diversity of lineages and the linear regressions (blue and pink dashed lines) of the lineage distributions generated by the two thresholds were compared. The area of the circle corresponds to the average SHM within the lineage. Black dotted line depicts y=x parity.

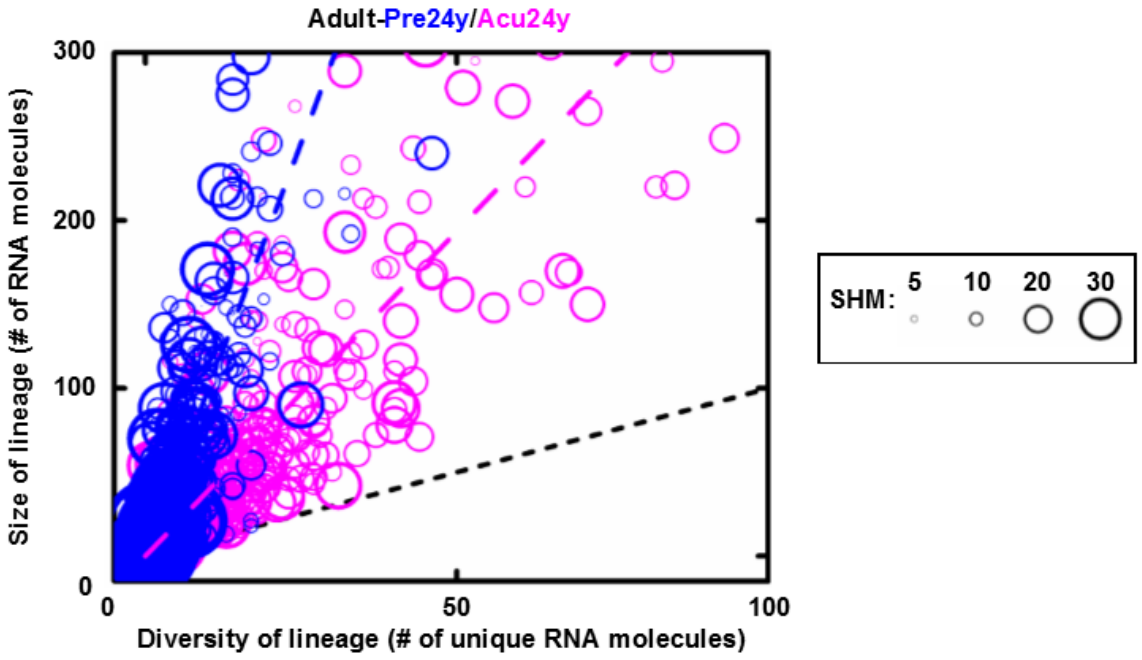

Figure B.6: Adult B cell lineage diversification. Size and diversity of B cell lineages between pre-malaria (blue) and acute malaria (pink) samples for a 24 year old adult malaria patient. Area of the circles corresponds to the average number of mutations within that lineage. Dashed lines represent the linear fit for pre- (blue) and acute (pink) lineages; black dotted line depicts y=x parity. Both axes were trimmed to be consistent with the main figures.

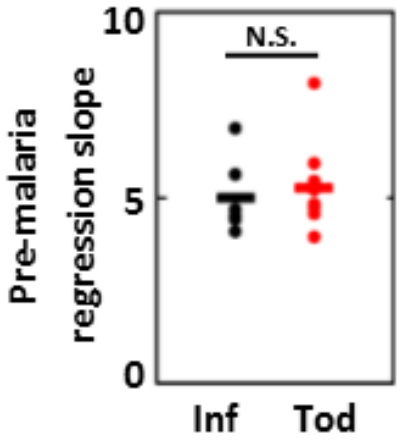

Figure B.7: Pre-malaria lineage diversification between infants and toddlers. Pre-malaria lineage size/diversity linear regression slopes (**Figure 3.9**, blue dashed lines) were compared between infants (black) and toddlers (red). N.S. indicates not significant by Mann Whitney U test, two-tailed. Bars indicate means.

Patient	Shared lineages	Unique memory B cell Sequences	Containing pre-malaria memory B cells
Inf1	29	N.A.	N.A.
Inf2	131	N.A.	N.A.
Inf3	215	N.A.	N.A.
Inf4	142	N.A.	N.A.
Inf5*	214	N.A.	N.A.
Inf6	83	N.A.	N.A.
Tod1*	308	3,423	149
Tod2	385	7,856	145
Tod3†	1230	6,023	926
Tod4	1194	5,073	209
Tod5	260	N.A.	N.A.
Tod6	346	6,363	111
Tod7†	472	4,771	161
Tod8	581	2,399	98
Tod9	414	2,534	135

Table B.6: Pre-malaria and acute malaria shared lineage count. The number of lineages containing sequences from both the pre-malaria and acute malaria timepoints. For malaria-experienced individuals with 10,000 FACS sorted pre-malaria memory B cells available, the number of unique memory B cell sequences and two-timepoint-shared lineages that contain sequences from the sorted memory B cells from the pre-malaria timepoint. N.A. indicates not applicable.*Same individual. † Same individual.

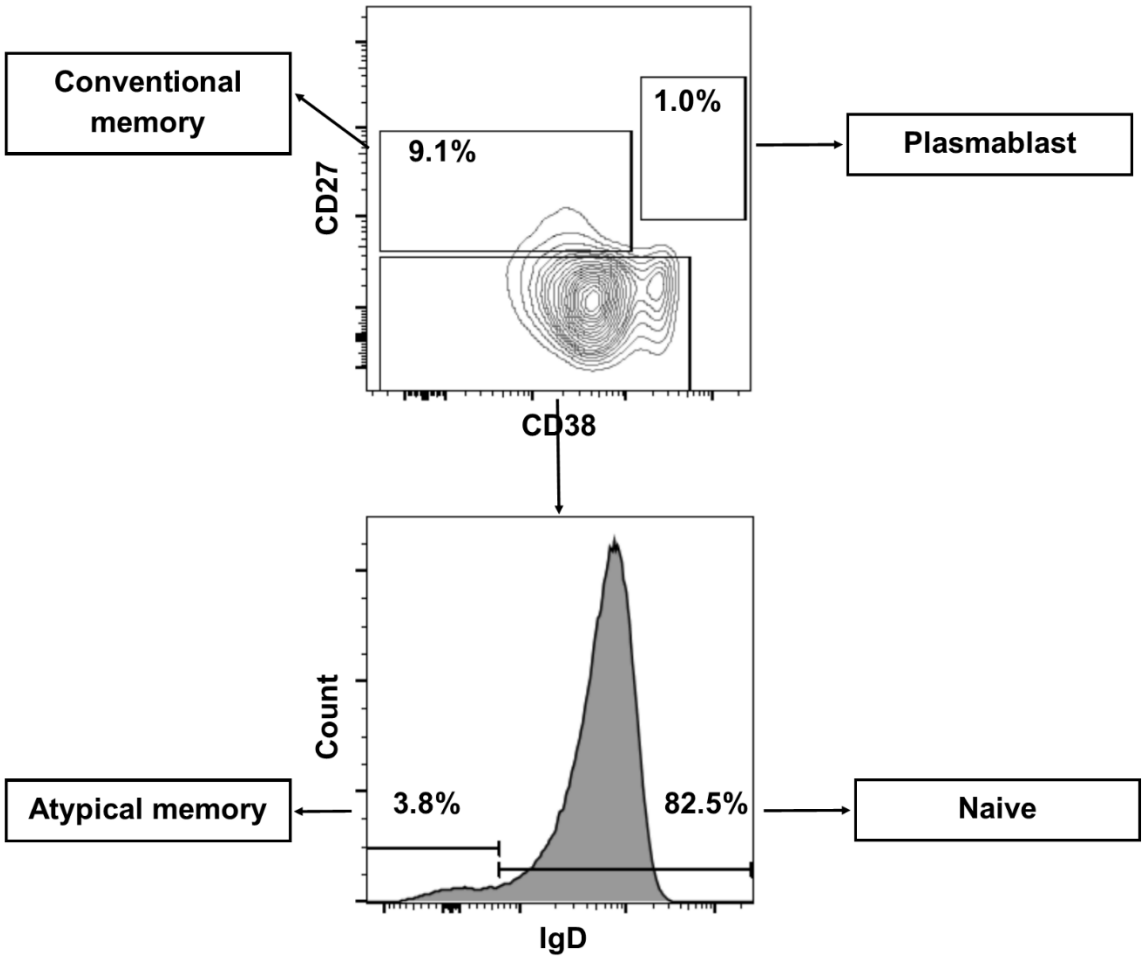

Figure B.8: Flow cytometry B cell gating and atypical memory percentage. B cells were first gated by scatter, then live, dump (CD4, CD8, CD14, CD56) negative, and then $CD19^{+}$. Conventional memory B cells ($CD20^{+}CD27^{+}$), plasmablasts ($CD27^{bright}CD38^{bright}$), and naïve B cells ($CD20^{+}CD27^{-}CD38^{low}$) were gated for further analysis. Atypical memory B cells ($CD20^{+}CD27^{-}CD38^{low}IgD^{-}$) make up a minor portion of the naïve-like B cells. Percentage of total B cells is displayed for each subpopulation.

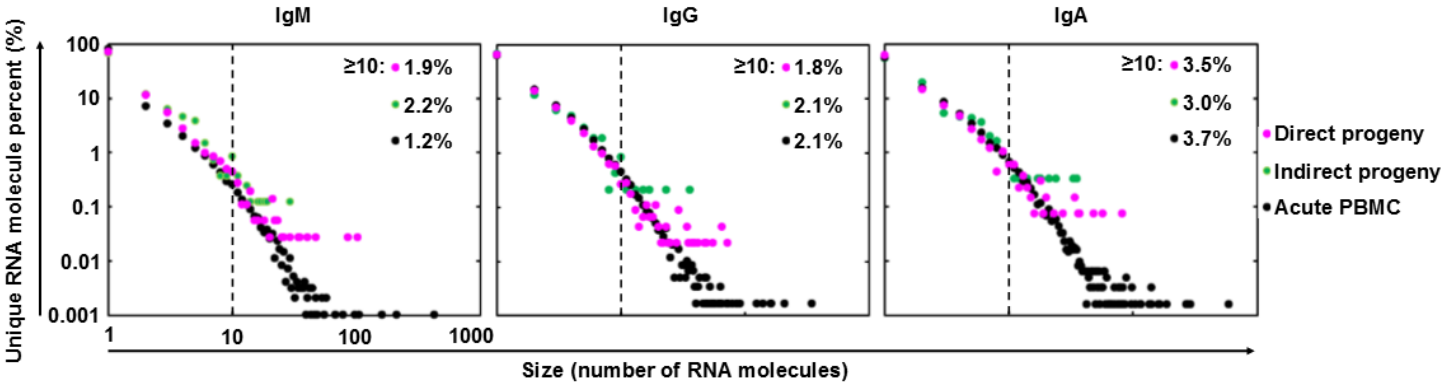

Figure B.9: Pre-malaria memory B cells' acute progeny RNA abundance. Shared lineages containing sequences from pre-malaria memory B cells and acute malaria PBMCs were formed as in **Figure 3.13** and **Figure 3.15**. Acute sequences from these lineages were classified as direct progeny (pink, corresponding to pink box in **Figure 3.14**) if they can be traced directly back to a pre-malaria memory B cell sequence or indirect progeny (green, corresponding to acute sequences in the same lineages as the dark blue slice in **Figure 3.14**) if they cannot (i.e. they stem from a separate branch in the lineage tree). The RNA abundance distribution for these sequences were split by isotype and compared to the bulk acute PBMCs (black) from the same individuals (N=8 toddlers, Tod5 was not included because there were insufficient cells for FACS sorting). Vertical dashed line indicates 10 RNA molecule cutoff, with the percentage of unique RNA molecules larger than this cutoff displayed in the top right corner of each panel.

a

	Germline	agtggtggttactactggagctggatccgccagcacccagggaagggcctggagtggattgggtacatctattacagtgg
0	Inf-Acu11m, IgG	....
1	Inf-Acu11m, IgA	....
2	Inf-Acu11m, IgA	....
3	Inf-Acu11m, IgG	....
4	Inf-Acu11m, IgG	...................a..
5	Inf-Acu11m, IgG	...................a..
6	Inf-Acu11m, IgG	....
7	Inf-Acu11m, IgG	......a..a................
8	Inf-Acu11m, IgA	....c...
9	Inf-Acu11m, IgA	....
10	Inf-Acu11m, IgG	....a......................c...
11	Inf-Acu11m, IgA	..................g...
	Germline	gagcacctactacaacccgtccctcaagagtcgagttaccatatcagtagacacgtctaagaaccagttctccctgaagc
0	Inf-Acu11m, IgG	....
1	Inf-Acu11m, IgA	....
2	Inf-Acu11m, IgA	....a.
3	Inf-Acu11m, IgG	....g....................
4	Inf-Acu11m, IgG	....
5	Inf-Acu11m, IgG	....
6	Inf-Acu11m, IgG	....g....................
7	Inf-Acu11m, IgG	....
8	Inf-Acu11m, IgA	...................g..
9	Inf-Acu11m, IgA	...............g..................c...t...
10	Inf-Acu11m, IgG	..............g...
11	Inf-Acu11m, IgA	...................g..
	Germline	tgagctctgtgactgccgcggacacggccgtgtattac..
0	Inf-Acu11m, IgG	....
1	Inf-Acu11m, IgA	....
2	Inf-Acu11m, IgA	....
3	Inf-Acu11m, IgG	....
4	Inf-Acu11m, IgG	....
5	Inf-Acu11m, IgG	....
6	Inf-Acu11m, IgG	...................a..
7	Inf-Acu11m, IgG	....
8	Inf-Acu11m, IgA	....
9	Inf-Acu11m, IgA	....
10	Inf-Acu11m, IgG	....
11	Inf-Acu11m, IgA	............g...........t.................................t..............t......
	Germline	
0	Inf-Acu11m, IgG	ggccagggaaccctggtcaccgtctcctca
1	Inf-Acu11m, IgA	....
2	Inf-Acu11m, IgA	....
3	Inf-Acu11m, IgG	....
4	Inf-Acu11m, IgG	....
5	Inf-Acu11m, IgG	....
6	Inf-Acu11m, IgG	....
7	Inf-Acu11m, IgG	....
8	Inf-Acu11m, IgA	....
9	Inf-Acu11m, IgA	....
10	Inf-Acu11m, IgG	....
11	Inf-Acu11m, IgA	....

b

	Germline	t g g g t g c g a c a g g c c c c t g g a c a a g g g c t t g a g t g g a t g g g a t g g a t c a g c g c t t a c a a t g g t a a c a c a a a c t a t g c a c a
0	Inf3-Acu11m, IgA	. . . . g . . . . a . . . . g . . . g . . . . . . a . . . .
1	Inf3-Acu11m, IgG	. . . . g . . . . g . a . . . . g . . . g . . . . . . a . . . .
2	Inf3-Acu11m, IgG	. . . . g . . . . a . . . . g . . . c . . . . . . a . . . .
3	Inf3-Acu11m, IgA	. . . . g . . . . a . . . . g . . . g . . . . . . a . . . .
4	Inf3-Acu11m, IgG	. . . . g . . . . a . . . . g . . . g . a c . . . a . . . .
5	Inf3-Acu11m, IgA	. . . . g . . . . a . . . . g . . . g . a . . . . a . . . . c . . . .
6	Inf3-Acu11m, IgG	. . . . g . . . . a . . . . g . . t . . a c . . . a . . . .
7	Inf3-Acu11m, IgA	. . . . g . . . . a . a . . a . . . g . . . . t . . . g . . . .

	Germline	g a a g c t c c a g g g c a g a g t c a c c a t g a c c a c a g a c a c a t c c a c g a g c a c a g c c t a c a t g g a g c t g a g g a g c c t g a g a t c t g
0	Inf3-Acu11m, IgA	. . . a . . . . g . . . . t . . . .
1	Inf3-Acu11m, IgG	. . . a . . . . g . . . . t . . . .
2	Inf3-Acu11m, IgG	. . . a . . . . c . . . . g . . . . t . . . .
3	Inf3-Acu11m, IgA	. . . a . . . . g . . . . . t . t . . . . . . t . . a . . . . a . .
4	Inf3-Acu11m, IgG	. . . a . . . . g . . . . g . . . . t . . . . . . t . . . .
5	Inf3-Acu11m, IgA	. . . a . . . . g . . . . t . . . . . a a . . . . . . a . . . .
6	Inf3-Acu11m, IgG	. . . a . . . . g . . . . g . . . . t . . . . . . t . . . .
7	Inf3-Acu11m, IgA	. . . a . . . . a . . . . c . . . . c . . . . .

	Germline	a c g a c a c g g c c g t g t a t t a c . . . .
0	Inf3-Acu11m, IgA	. . . .
1	Inf3-Acu11m, IgG	. . . . t . . . .
2	Inf3-Acu11m, IgG	. . . . c . . . .
3	Inf3-Acu11m, IgA	. . . .
4	Inf3-Acu11m, IgG	. . . .
5	Inf3-Acu11m, IgA	. . . . a . . . .
6	Inf3-Acu11m, IgG	. . . . t . . . .
7	Inf3-Acu11m, IgA	. . . .

	Germline	g g c c a g g g a a c c c t g g t c a c c g t c t c c t c a
0	Inf3-Acu11m, IgA	. . . .
1	Inf3-Acu11m, IgG	. . . .
2	Inf3-Acu11m, IgG	. . . . g . . . .
3	Inf3-Acu11m, IgA	. . . .
4	Inf3-Acu11m, IgG	. . . .
5	Inf3-Acu11m, IgA	. . . .
6	Inf3-Acu11m, IgG	. . . .
7	Inf3-Acu11m, IgA	. . . .

c

	Germline	cactgggtccgccaggcttccgggaaagggctggagtgggttggccgtattagaagcaaagctaacagttacgcgacagc
0	Tod4-Pre29m, IgM	....a........
1	Tod4-Acu32m, IgM	....c................a............
2	Tod4-Acu32m, IgG	....c................a............
3	Tod4-Acu32m, IgA	....c................a............
4	Tod4-Acu32m, IgA	....c................a............
5	Tod4-Acu32m, IgA	....c................a............
6	Tod4-Acu32m, IgA	....c........c................a............
7	Tod4-Acu32m, IgA	....c.........a......a............
8	Tod4-Acu32m, IgA	....c................a............
9	Tod4-Acu32m, IgA	....c................a............
	Germline	atatgctgcgtcggtgaaaggcaggttcaccatctccagagatgattcaaagaacacggcgtatctgcaaatgaacagcc
0	Tod4-Pre29m, IgM	..tc.............c...g......
1	Tod4-Acu32m, IgM	..tc............g..t....................
2	Tod4-Acu32m, IgG	..tc............c..t....................
3	Tod4-Acu32m, IgA	..tc............c..t..t.................
4	Tod4-Acu32m, IgA	..tc............c..t..t.................
5	Tod4-Acu32m, IgA	..tc............c..t..t.................
6	Tod4-Acu32m, IgA	..tc............c..t..t.................
7	Tod4-Acu32m, IgA	..tc............c..t....................
8	Tod4-Acu32m, IgA	..tc...........ac............t.............................t....................
9	Tod4-Acu32m, IgA	..tc...........ac............t.............................t.c..................
	Germline	tgaaaaccgaggacacggccgtgtattac...
0	Tod4-Pre29m, IgM	....
1	Tod4-Acu32m, IgM	....
2	Tod4-Acu32m, IgG	....................a...........g......................g........................
3	Tod4-Acu32m, IgA	....................a...........g...
4	Tod4-Acu32m, IgA	....................a.......t...g...
5	Tod4-Acu32m, IgA	....................a...........g......................g........................
6	Tod4-Acu32m, IgA	....................a...........g......................g........................
7	Tod4-Acu32m, IgA	....................a...........g......................g........................
8	Tod4-Acu32m, IgA	..........a.........a.c.........g.......................g.......................
9	Tod4-Acu32m, IgA	..........a.........a.c.........g.......................g.......................
	Germline	ggccagggaaccctggtcaccgtctcctca
0	Tod4-Pre29m, IgM	....
1	Tod4-Acu32m, IgM	....
2	Tod4-Acu32m, IgG	....
3	Tod4-Acu32m, IgA	....
4	Tod4-Acu32m, IgA	....
5	Tod4-Acu32m, IgA	....
6	Tod4-Acu32m, IgA	....
7	Tod4-Acu32m, IgA	..............a...............
8	Tod4-Acu32m, IgA	....
9	Tod4-Acu32m, IgA	....

Figure B.10 Sequence alignment for illustrated lineages. The CDR3 region has been highlighted in yellow. The top row displays the IMGT germline allele sequence, and dashes indicate where the sequences are identical to the germline. (**a**) Corresponds to the lineage in **Figure 3.10a**, (**b**) corresponds to the lineage in **Figure 3.10b** and (**c**) corresponds to the lineage in **Figure 3.13**.

Appendix C – Chapter 4 Supplementary Information

Target	Forward Primer Sequence	Reverse Primer Sequence	PCR Stage
IGHV1-8*02 (G234T)	GGGCTGAGGTGAAGAAGC	CCTCTCGCACAGTAATACACG	1st PCR
IGHV3-30*02 (T201C)	GTGCAGCTGGTGGAGTC	CTTTCGCACAGTAATACACAGC	
IGHV4-61*01 (C93T_C136G_A138C)	GACTGGTGAAGCCTTCGG	TCTCTCGCACAGTAATACACG	
IGHV4-59*01 (T109C)	GCCCAGGACTGGTGAAG	TCTCTCGCACAGTAATACACG	
IGHV1-69*01 (G163A)	GGGCTGAGGTGAAGAAGC	TCTCTCGCACAGTAATACACG	
IGHV4-31*02 (C198T)	GCCCAGGACTGGTGAAG	TCTCTCGCACAGTAATACACG	
IGHV1-8*02 (G234T)	GACGTGTGCTCTTCCGATCT GGGCCTCAGTGAAGGTCT	ACACTCTTTCCCTACACGACGCTCTTC CGATCT TCAGATCTCAGGCTGCTCA	Nested PCR
IGHV3-30*02 (T201C)	GACGTGTGCTCTTCCGATCT GTCCCTGAGACTCTCCTGT	ACACTCTTTCCCTACACGACGCTCTTC CGATCT AGCTCTCAGGCTGTTCATTT	
IGHV4-61*01 (C93T_C136G_A138C)	GACGTGTGCTCTTCCGATCT CTCACCTGCACTGTCTCTG	ACACTCTTTCCCTACACGACGCTCTTC CGATCT GTCACAGAGCTCAGCTTCA	
IGHV4-59*01 (T109C)	GACGTGTGCTCTTCCGATCT GACCCTGTCCCTCACCT	ACACTCTTTCCCTACACGACGCTCTTC CGATCT TCACAGAGCTCAGCTTCA	
IGHV1-69*01 (G163A)	GACGTGTGCTCTTCCGATCT TCCTCGGTGAAGGTCTCC	ACACTCTTTCCCTACACGACGCTCTTC CGATCT GCTGCTCAGCTCCATGT	
IGHV4-31*02 (C198T)	GACGTGTGCTCTTCCGATCT CCCTGTCCCTCACCTGTA	ACACTCTTTCCCTACACGACGCTCTTC CGATCT GTCACAGAGCTCAGCTTCA	
Illumina Adaptors	CAAGCAGAAGACGGCATAC GAGATAANNNNNN GTGACTGGAGTTCAGACGT GTGCTCTTCCGATCT	AATGATACGGCGACCACCGAGATCT ACACTCTTTCCCTACACGAC	Adaptor Extension

Table C.1: gDNA validation primer list. Sequences in red indicate common partial Illumina adaptors; NNNNNN in blue indicates fixed library indexes used to pool multiple libraries into a single run.

113	CD57	HCD57	Biolegend
141	CD3	UCTH1	BD
142	CD5	UCHT2	Biolegend
143	CD8	SK1	Biolegend
144	CD4	SK3	Biolegend
145	CD19	HIB19	Biolegend
146	IgD	IA6-2	Biolegend
147	IFNg	4S.B3	eBioscience
148	CCR6	G034E3	Biolegend
149	CD14	M5E2	Biolegend
150	CD69	FN50	Biolegend
151	Granzyme A	CB9	Biolegend
152	TNFa	MAb11	Biolegend
153	CD45RO	UCHL1	Biolegend
154	CD27	LG.7F9	eBioscience
155	TCRab	T10B9.1A-31	BD
156	CCR5	J418F1	Biolegend
157	Ki67	B56	BD
158	BLYS	1D6	Biolegend
159	BCL6	K112-91	BD
160	IL4	8D4-8	BD
161	CD25	M-A251	BD
162	Eomes	wd1928	eBioscience
163	ICOS	c398-4A	Biolegend
164	IL5	TRFK5	Biolegend
165	Foxp3	PCH101	eBioscience
166	CD95	DX2	Biolegend
167	CD45RA	HI100	eBioscience
168	CCR7	G043H7	Biolegend
169	CXCR5	RF8B2	BD
170	Tbet	4B10	Biolegend
171	PD-1	EH12.2H7	Biolegend
172	IL-21	3A3-N2	Biolegend
173	IL2	MQ1-17H12	eBioscience
174	CD24	ML5	Biolegend
175	CXCR3	G025H7	Biolegend
176	CD38	HIT2	Biolegend

Table D.1: CyTOF antibody staining panel. Table produced by L.F.S.

ID	Study groups	Gender	Age	Site	CD4 Tc Cells/ul	CD4:CD8 Ratio	pVL (RNA copies/ml)	ART
H1	HIV	F	41	cervical	7	0.06	230948	2 months
H2	HIV	M	29	cervical	157	0.19	182520	No
H3	HIV	M	19	cervical	157	0.15	63	3 months
H4	HIV	M	23	cervical	199	0.29	3034	No
H5	HIV	F	20	cervical	212	0.28	136190	No
H6	HIV	M	22	cervical	223	0.08	280700	1 day
H7	HIV	M	29	cervical	241	0.22	447952	No
H8	HIV	M	33	cervical	254	0.32	75043	No
H9	HIV	M	19	cervical	258	0.41	35847	No
H10	HIV	M	37	cervical	276	0.17	199152	No
H11	HIV	F	40	cervical	279	0.3	111854	No
H12	HIV	M	29	cervical	295	0.42	71	36 months
H13	HIV	M	29	cervical	367	0.35	undetectable <40	7 months
H14	HIV	M	38	cervical	387	0.4	71004	No
H15	HIV	M	26	cervical	399	1.03	9610	No
H16	HIV	M	29	cervical	462	0.36	56009	No
H17	HIV	M	23	cervical	473	0.62	208656	No
H18	HIV	M	18	cervical	504	0.71	undetectable <40	No
H19	HIV	M	22	cervical	538	0.32	9511	No
H20	HIV	M	28	cervical	564	0.4	16647	No
H21	HIV	M	21	cervical	573	0.6	17868	No
H22	HIV	M	40	cervical	668	0.57	17247	No
H23	HIV	M	20	cervical	743	0.58	83801	No
H24	HIV	M	28	cervical	1086	0.77	3964	No
H25	HIV	M	22	cervical	1136	0.72	45395	No
C1	HC	F	59	mesenteric				
C2	HC	F	58	mesenteric				
C3	HC	NA	NA	iliac				
C4	HC	NA	NA	iliac				
C5	HC	NA	NA	iliac				
C6	HC	NA	NA	iliac				
C7	HC	NA	NA	cervical				
IBD1	Crohn's	F	30	mesenteric				
IBD2	UC	M	29	mesenteric				
IBD3	Crohn's	M	54	mesenteric				
IBD4	UC	M	37	mesenteric				

NA= not available

Table D.2: Clinical characteristics and demographic information of LN samples. Table produced by L.F.S.

	Naïve		Memory		GC T_{FH}	
Subject	Cells	TCR transcripts	Cells	TCR transcripts	Cells	TCR transcripts
H2	10,000	14,420	10,000	6,315	15,000	33,904
H3	10,000	6,750	10,000	8,945	10,000	7,954
H8	10,000	13,225	10,000	5,992	10,000	26,374
H10	10,000	4,314	10,000	9,033	1,464	2,498
H14	10,000	5,822	10,000	6,254	10,000	11,197
H17	10,000	18,376	10,000	6,533	10,995	37,129
H21	10,000	10,404	10,000	7,503	10,227	23,673
H24	10,000	5,488	10,000	2,821	10,000	9,220

Table D.3: TCR repertoire sequencing cell and transcript counts.

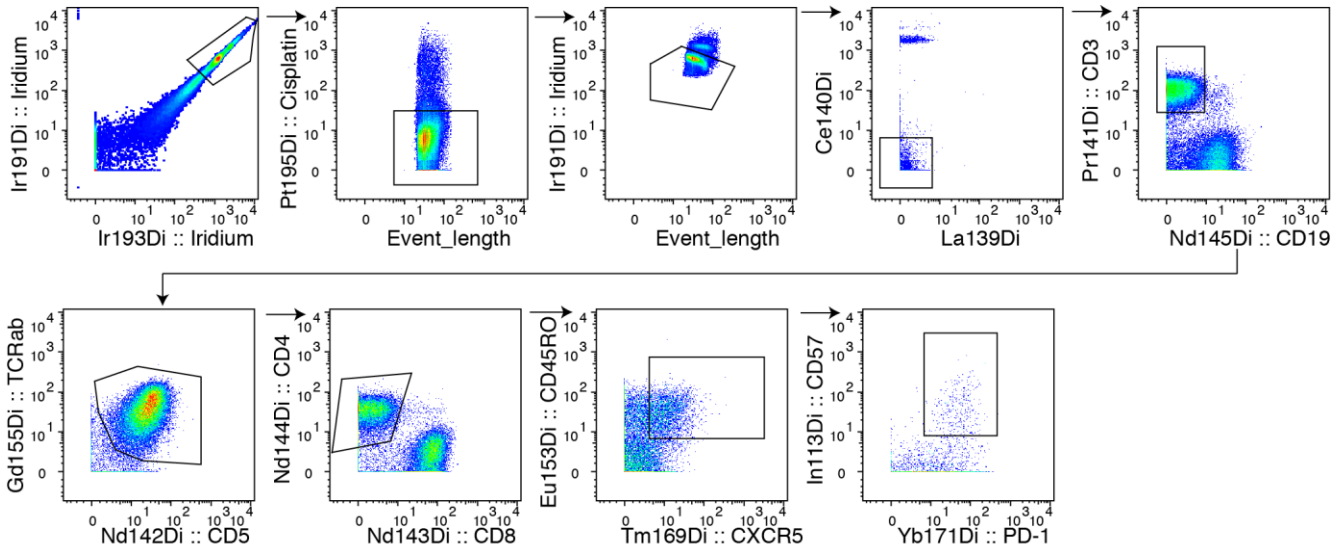

Figure D.1: Identification of GC T_{FH} cells from LN samples. Representative gating to identify GC T_{FH} cells using data from an HIV^+ sample. Cryopreserved LN cells were stimulated with PMA and ionomycin in the presence of Brefeldin A and monensin, stained with a panel of 37 surface and intracellular markers, and analyzed by mass cytometry. After exclusion of dead cells ($Cisplatin^+$), doublets (by event length), beads ($Cd140^+$), and elimination of background signal in an empty gate (La139), lineage gating was performed to identify $CD3^+CD4^+TCR\alpha\beta^+$ T cells. GC T_{FH} cells were identified as the subset of $CXCR5^+$ $CD45RO^+$ $CD4^+$ T cells that express CD57 and PD-1. Data and figure produced by L.F.S.

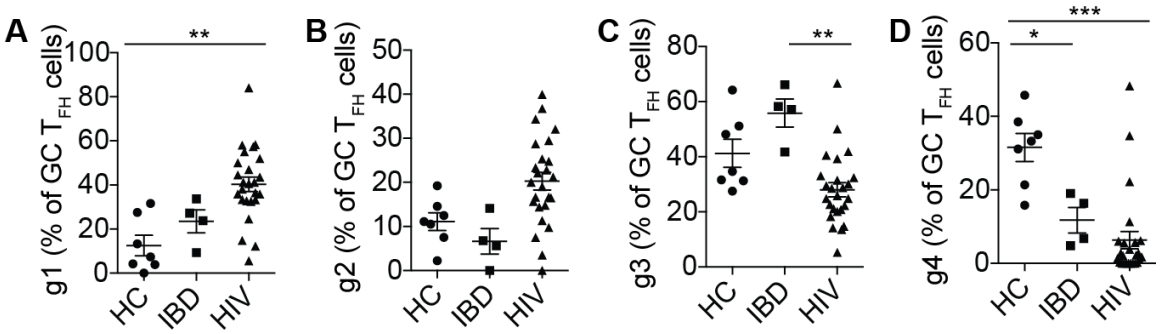

Figure D.2: Phenotypic group distribution is different between HC, IBD, and HIV samples. Frequency of g1 (**A**), g2 (**B**), g3 (**C**), or g4 (**D**) in HC, IBD, and HIV samples. Statistical significance was analyzed using two-tailed Student's t-test for pair-wise comparison. Significance level is set at $p < 0.0167$ to correct for three-way comparison. * *P < 0.0167,* ** *P < 0.00167,* *** *P < 0.000167*. Data and figure produced by L.F.S.

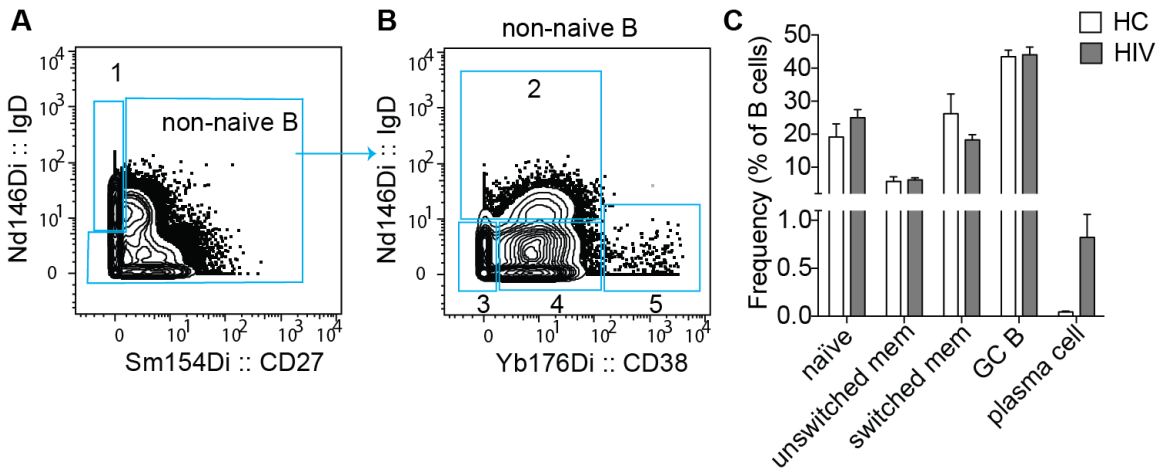

Figure D.3: Identification of B cell subsets. (**A**) Plots showing gating strategy to identify naïve B cells ($IgD^{+}CD27^{-}$, 1). (**B**) Non-naïve subsets are divided based on IgD and CD38 expression into unswitched memory B cells (IgD^{+}, 2), switched memory B cells ($IgD^{-}CD38^{-}$, 3), GC B cells ($IgD^{-}CD38^{+}$, 4), and plasma cells ($IgD^{-}CD38^{high}$, 5). (**C**) Bar-graph showing the distribution of B cell subsets in HIV^{+} and HC LNs. Differences between HIV and HC were not statistically significant by Student's t-test. Data and figure produced by L.F.S.

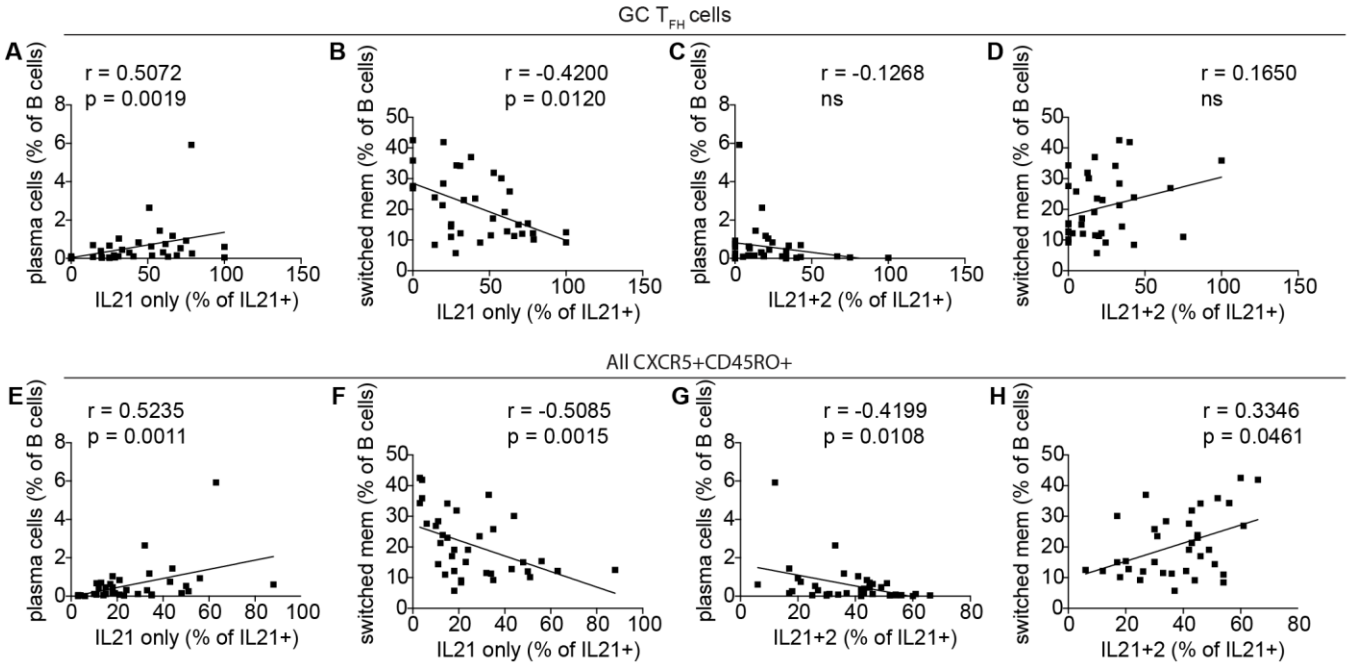

Figure D.4: IL-21 poly-functionality is associated with distinct B cell phenotypes. (**A-D**) Correlation between plasma cells or switched B cells with IL-21 only GC T_{FH} cells (**A,B**). An inverse trend is observed for IL21+2 GC T_{FH} cells (**C,D**). (**E-H**) Correlation between IL-21 only and IL21+2 $CXCR5^{+}CD45RO^{+}CD4^{+}$ T cells with plasma cells and switched B cells as in **A-D**. Spearman rank correlation and least squares fit regression were applied to measure the degree of association. Data and figure produced by L.F.S.

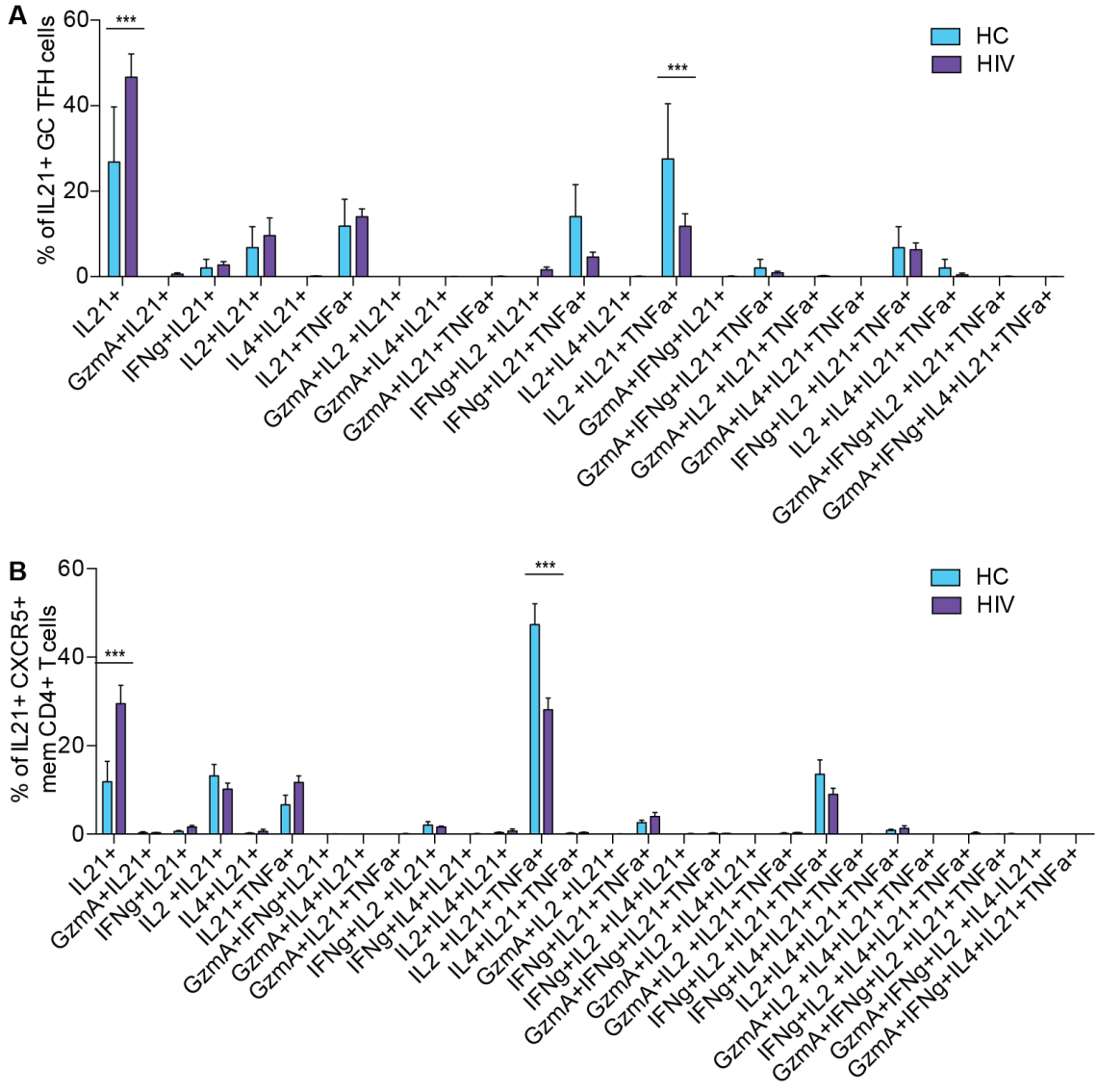

Figure D.5: Frequency of IL-21-producing subsets in T_{FH} cells in HIV and HC samples. (**A,B**) Bar graph shows all cytokine-positive IL-21-producing subsets in combination with other effector molecules (IL-2, IFN-γ, TNF-α, IL-4, or granzyme A). These subsets were generated by applying Boolean combination gating on IL-21$^+$ GC T_{FH} (**A**) or CXCR5$^+$ memory CD4+ T cells (**B**). Statistical significance was analyzed using Student's t-test for pair-wise comparison. Multiple comparisons are corrected using Holm-Sidak method, with alpha=5%. *** $P < 0.0005$. Data and figure produced by L.F.S.

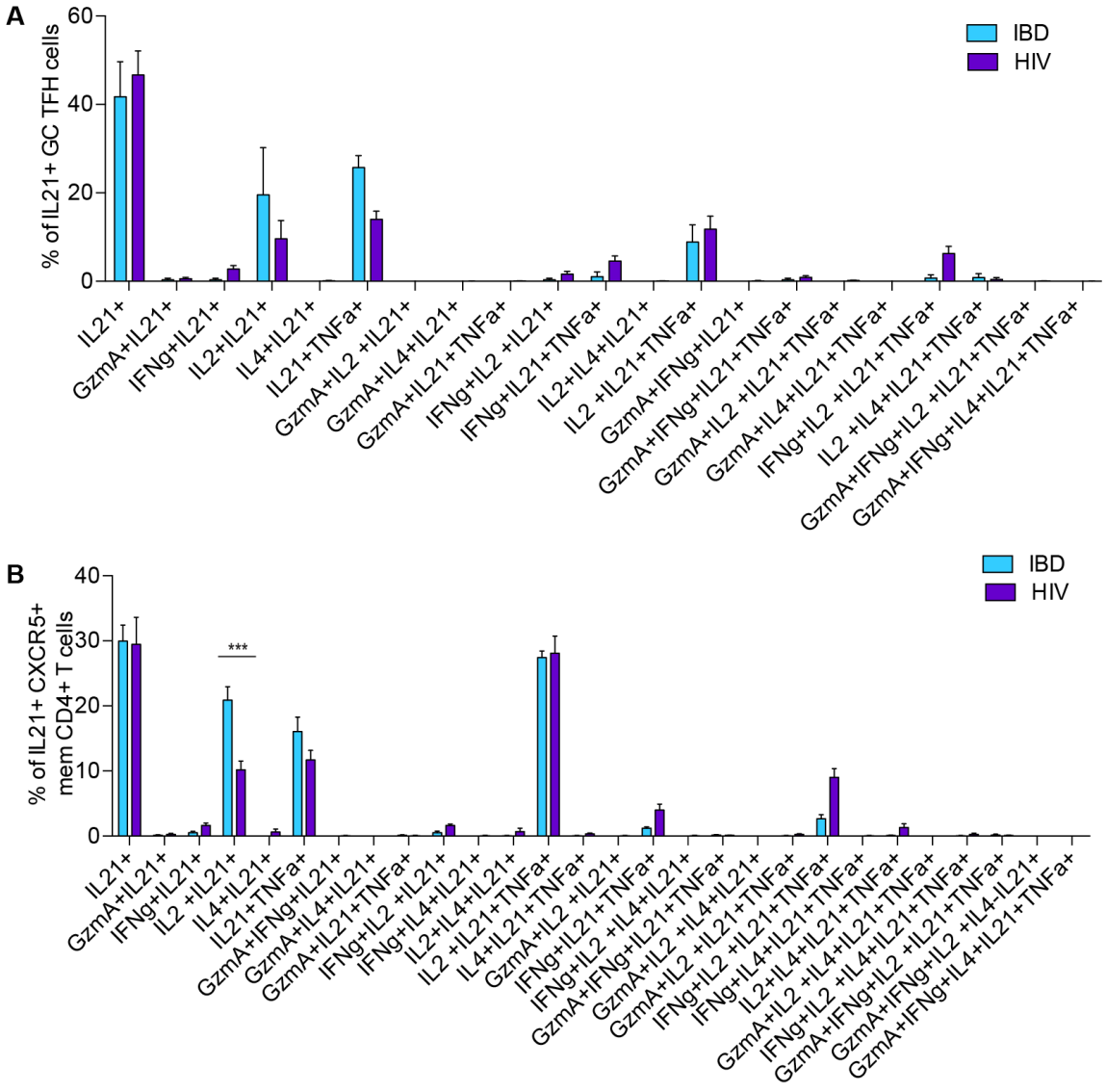

Figure D.6: Frequency of IL-21-producing subsets in T_{FH} cells in HIV and IBD samples. (**A-B**) Bar graph shows all cytokine positive IL-21 producing subsets in combination with other effector molecules (IL-2, IFN-γ, TNF-α, IL-4, or granzyme A). These subsets were generated as in **Figure D.5**. Statistical significance was analyzed using Student's t-test for pair-wise comparison. Multiple comparisons are corrected using Holm-Sidak method, with alpha=5%. *** $P < 0.0005$. Data and figure produced by L.F.S.

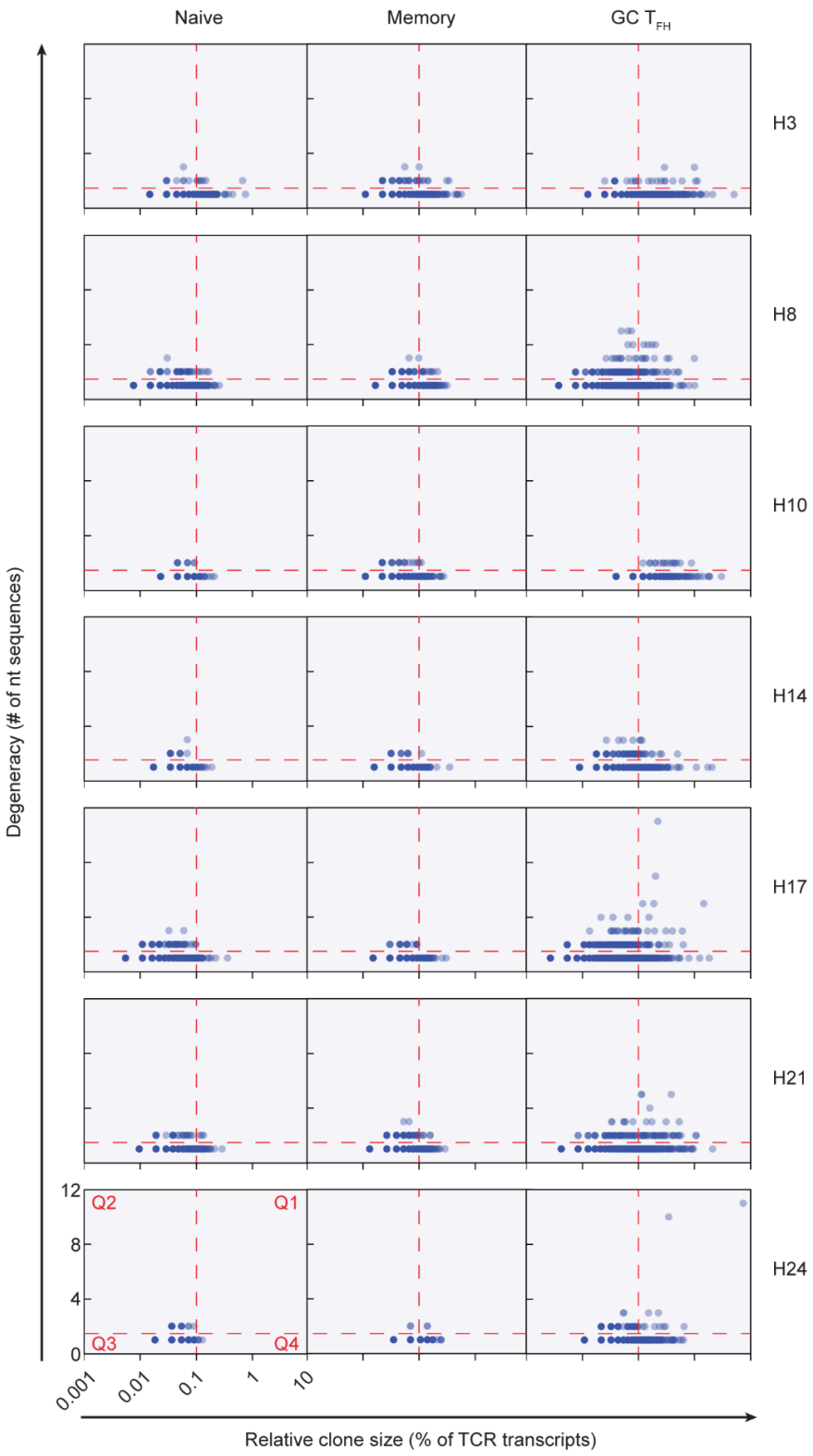
Naive
Memory
GC T FH
H3
H8
H10
H14
H17
H21
H24
Q2
Q1
Q3
Q4
12
8
4
0
0.001
0.01
0.1
1
10
Degeneracy (# of nt sequences)
Relative clone size (% of TCR transcripts)

Figure D.7

Figure D.7: Antigen-driven clonal selection signature in GC T_{FH} cells of HIV^+ LNs. Coding degeneracy level (number of unique TCR nucleotide (nt) sequences encoding a common CDR3 amino acid (aa) sequence) of each CDR3 aa sequence is plotted against their frequency (measured as % of total TCR transcript) in naïve, memory, and GC T_{FH} cells. Each dot is a unique CDR3 aa sequence. Red dashed lines indicate cutoffs for degenerate (2 or more nt sequences coding for the same aa sequence, horizontal) and expanded (0.1% or more of TCR transcripts, vertical) clones. Each panel is broken into 4 quadrants: Q1: degenerate-abundant clones; Q2: degenerate-rare clones; Q3: nondegenerate-rare clones; Q4: nondegenerate-abundant clones. See also **Figure 5.6A**.

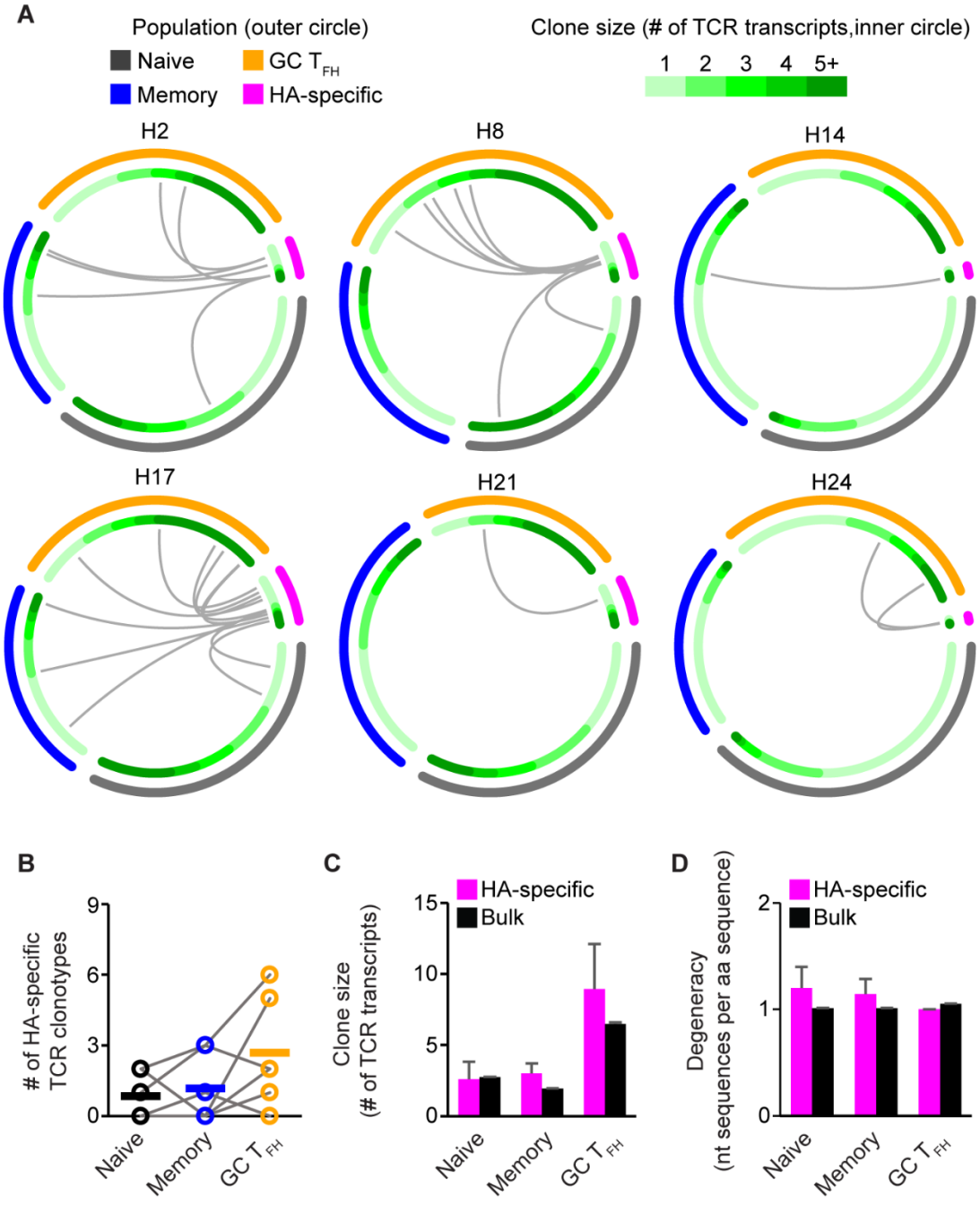
A
Population (outer circle)
Naive
GC T FH
Memory
HA-specific
Clone size (# of TCR transcripts,inner circle)
1 2 3 4 5+
H2
H8
H14
H17
H21
H24
B
# of HA-specific TCR clonotypes
9
6
3
0
Naive
Memory
GC T FH
C
Clone size (# of TCR transcripts)
HA-specific
Bulk
15
10
5
0
Naive
Memory
GC T FH
D
Degeneracy (nt sequences per aa sequence)
HA-specific
Bulk
2
1
0
Naive
Memory
GC T FH

Figure D.8

Figure D.8: HA-specific $CD4^+$ T cells within HIV^+ LNs lack clonal expansion and selection signatures. (**A**) HA-specific TCR clones overlap with HIV^+ LN $CD4^+$ T cell populations. Each thin slice of the arc represents a unique TCR sequence, ordered by the clone size (darker green for larger clones, inner circle). Grey curves indicate HA-specific TCR clones (nt sequences) found in naïve (black, outer circle), memory (blue, outer circle), and GC T_{FH} (orange, outer circle) populations. (**B**) Number of HA-specific TCR clones observed in naïve (black), memory (blue), and GC T_{FH} (orange) populations. Grey lines connect samples from the same patient. Bars indicate means. (**C**) Mean clone size of HA-specific clones (magenta) and clones of unknown specificity (black) from naïve, memory, and GC T_{FH} populations. (**D**) Degeneracy, or the number of distinct nt sequences per CDR3 aa sequence, of HA-specific clones (magenta) and clones of unknown specificity (black) from naïve, memory, and GC T_{FH} populations. Data from all 6 subjects were aggregated for C and D. Error bars indicate SEM.

Primer Target	Sequence	PCR Step
TCRb Constant	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN NNN NNN GAC CTC GGG TGG GAA CAC	Reverse Transcription
TRBV1	GAC GTG TGC TCT TCC GAT CTC TGA CAG CTC TCG CTT ATA CCT TCA	1st PCR, forward
TRBV2	GAC GTG TGC TCT TCC GAT CTG CCT GAT GGA TCA AAT TTC ACT CTG	
TRBV3	GAC GTG TGC TCT TCC GAT CTA ATG AAA CAG TTC CAA ATC GMT TCT	
TRBV4	GAC GTG TGC TCT TCC GAT CTC CAA GTC GCT TCT CAC CTG AAT	
TRBV5-1	GAC GTG TGC TCT TCC GAT CTC GCC AGT TCT CTA ACT CTC GCT CT	
TRBV5-2	GAC GTG TGC TCT TCC GAT CTT TAC TGA GTC AAA CAC GGA GCT AGG	
TRBV5-3	GAC GTG TGC TCT TCC GAT CTC TCT GAG ATG AAT GTG AGT GCC TTG	
TRBV5-4/5/6/7/8	GAC GTG TGC TCT TCC GAT CTC TGA GCT GAA TGT GAA CGC CTT G	
TRBV6-1	GAC GTG TGC TCT TCC GAT CTT CTC CAG ATT AAA CAA ACG GGA GTT	
TRBV6-2/3	GAC GTG TGC TCT TCC GAT CTC TGA TGG CTA CAA TGT CTC CAG ATT	
TRBV6-4	GAC GTG TGC TCT TCC GAT CTA GTG TCT CCA GAG CAA ACA CAG ATG	
TRBV6-5/6/7	GAC GTG TGC TCT TCC GAT CTG TCT CCA GAT CAA MCA CAG AGG ATT	
TRBV6-8/9	GAC GTG TGC TCT TCC GAT CTA AAC ACA GAG GAT TTC CCR CTC AG	
TRBV7-1	GAC GTG TGC TCT TCC GAT CTG TCT GAG GGA TCC ATC TCC ACT C	
TRBV7-2	GAC GTG TGC TCT TCC GAT CTT CGC TTC TCT GCA GAG AGG ACT GG	
TRBV7-3	GAC GTG TGC TCT TCC GAT CTC TGA GGG ATC CGT CTC TAC TCT GAA	
TRBV7-4/8	GAC GTG TGC TCT TCC GAT CTC TGA GRG ATC CGT CTC CAC TCT G	
TRBV7-5	GAC GTG TGC TCT TCC GAT CTG GTC TGA GGA TCT TTC TCC ACC T	
TRBV7-6/7	GAC GTG TGC TCT TCC GAT CTG AGG GAT CCA TCT CCA CTC TGA C	
TRBV7-9	GAC GTG TGC TCT TCC GAT CTC TGC AGA GAG GCC TAA GGG ATC T	
TRBV8-1	GAC GTG TGC TCT TCC GAT CTA AGC TCA AGC ATT TTC CCT CAA C	
TRBV8-2	GAC GTG TGC TCT TCC GAT CTA TGT CAC AGA GGG GTA CTG TGT TTC	
TRBV9	GAC GTG TGC TCT TCC GAT CTA CAG TTC CCT GAC TTG CAC TCT G	
TRBV10-1/3	GAC GTG TGC TCT TCC GAT CTA CAA AGG AGA AGT CTC AGA TGG CTA	
TRBV10-2	GAC GTG TGC TCT TCC GAT CTT GTC TCC AGA TCC AAG ACA GAG AA	
TRBV11	GAC GTG TGC TCT TCC GAT CTC TGC AGA GAG GCT CAA AGG AGT AG	
TRBV12-1/2	GAC GTG TGC TCT TCC GAT CTA TCA TTC TCY ACT CTG AGG ATC CAR	
TRVB12-3/4/5	GAC GTG TGC TCT TCC GAT CTA CTC TGA RGA TCC AGC CCT CAG AAC	
TRBV13	GAC GTG TGC TCT TCC GAT CTC AGC TCA ACA GTT CAG TGA CTA TCA T	
TRBV14	GAC GTG TGC TCT TCC GAT CTG AAA GGA CTG GAG GGA CGT ATT CTA	
TRBV15	GAC GTG TGC TCT TCC GAT CTG CCG AAC ACT TCT TTC TGC TTT CT	
TRBV16	GAC GTG TGC TCT TCC GAT CTA TTT TCA GCT AAG TGC CTC CCA AAT	
TRBV17	GAC GTG TGC TCT TCC GAT CTC ACA GCT GAA AGA CCT AAC GGA AC	
TRBV18	GAC GTG TGC TCT TCC GAT CTA TTT TCT GCT GAA TTT CCC AAA GAG	
TRBV19	GAC GTG TGC TCT TCC GAT CTG TCT CTC GGG AGA AGA AGG AAT C	
TRBV20-1	GAC GTG TGC TCT TCC GAT CTG ACA AGT TTC TCA TCA ACC ATG CAA	
TRBV21-1	GAC GTG TGC TCT TCC GAT CTC AAT GCT CCA AAA ACT CAT CCT GT	
TRBV22-1	GAC GTG TGC TCT TCC GAT CTA GGA GAA GGG GCT ATT TCT TCT CAG	
TRBV23-1	GAC GTG TGC TCT TCC GAT CTA TTC TCA TCT CAA TGC CCC AAG AAC	
TRBV24-1	GAC GTG TGC TCT TCC GAT CTG ACA GGC ACA GGC TAA ATT CTC C	
TRBV25-1	GAC GTG TGC TCT TCC GAT CTA GTC TCC AGA ATA AGG ACG GAG CAT	
TRBV26	GAC GTG TGC TCT TCC GAT CTC TCT GAG GGG TAT CAT GTT TCT TGA	
TRBV27	GAC GTG TGC TCT TCC GAT CTC AAA GTC TCT CGA AAA GAG AAG AGG A	
TRBV28	GAC GTG TGC TCT TCC GAT CTA AGA AGG AGC GCT TCT CCC TGA TT	
TRBV29-1	GAC GTG TGC TCT TCC GAT CTC GCC CAA ACC TAA CAT TCT CAA	
TRBV30	GAC GTG TGC TCT TCC GAT CTC CAG AAT CTC TCA GCC TCC AGA C	
ILLUPE1adaptor_short	ACA CTC TTT CCC TAC ACG AC	1st PCR, reverse
ILLUPE2adaptor_full	CAA GCA GAA GAC GGC ATA CGA GAT AA NNN NNN GTG ACT GGA GTT CAG ACG TG	2nd PCR, forward
ILLUPE1adaptor_full	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC	2nd PCR, reverse

Table D.4: TCRβ Sequencing Primers. Red Ns indicate 12N random molecular identified (MID). Blue Ns indicate fixed Illumina i7 indexes used for pooling multiple libraries for a single run.

References

1. Efficacy and safety of the RTS,S/AS01 malaria vaccine during 18 months after vaccination: a phase 3 randomized, controlled trial in children and young infants at 11 African sites. *PLoS medicine* **11**, e1001685 (2014).

2. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet* **386**, 31-45 (2015).

3. Adkins B, Leclerc C, Marshall-Clarke S. Neonatal adaptive immunity comes of age. *Nature reviews Immunology* **4**, 553-564 (2004).

4. Ahmed R, Gray D. Immunological memory and protective immunity: understanding their relation. *Science* **272**, 54-60 (1996).

5. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* **882**, 569-604 (2012).

6. Barouch DH*, et al.* Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* **503**, 224-228 (2013).

7. Baumjohann D*, et al.* Persistent antigen and germinal center B cells sustain T follicular helper cell responses and phenotype. *Immunity* **38**, 596-605 (2013).

8. Biswas S, Saxena QB, Roy A, Kabilan L. Naturally occurring plasmodium-specific IgA antibody in humans from a malaria endemic area. *Journal of Biosciences* **20**, 453-460 (1995).

9. Bolotin DA*, et al.* Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *European journal of immunology* **42**, 3073-3083 (2012).

10. Boyd SD*, et al.* Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* **184**, 6986-6992 (2010).

11. Boyd SD*, et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science*

translational medicine **1**, 12ra23 (2009).

12. Chang B, Casali P. The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. *Immunology today* **15**, 367-373 (1994).

13. Chen K, Gogu V, Wu D, Jiang N. COLT: Constrained Lineage Tree Generation from Sequence Data. *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, (2016).

14. Corcoran MM*, et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications* **7**, 13642 (2016).

15. Corti D*, et al.* Protective monotherapy against lethal Ebola virus infection by a potently neutralizing antibody. *Science* **351**, 1339-1342 (2016).

16. Crotty S. Follicular helper CD4 T cells (TFH). *Annual review of immunology* **29**, 621-663 (2011).

17. Crotty S. T follicular helper cell differentiation, function, and roles in disease. *Immunity* **41**, 529-542 (2014).

18. Crum-Cianflone NF*, et al.* Durability of antibody responses after receipt of the monovalent 2009 pandemic influenza A (H1N1) vaccine among HIV-infected and HIV-uninfected adults. *Vaccine* **29**, 3183-3191 (2011).

19. Davila ML*, et al.* Efficacy and toxicity management of 19-28z CAR T cell therapy in B cell acute lymphoblastic leukemia. *Science translational medicine* **6**, 224ra225 (2014).

20. De Milito A*, et al.* Mechanisms of hypergammaglobulinemia and impaired antigen-specific humoral immunity in HIV-1 infection. *Blood* **103**, 2180-2186 (2004).

21. De Silva NS, Klein U. Dynamics of B cells in germinal centres. *Nature reviews Immunology* **15**, 137-148 (2015).

22. DeKosky BJ*, et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature biotechnology* **31**, 166-169 (2013).

23. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic

hypermutation. *Annual review of biochemistry* **76**, 1-22 (2007).

24. Dogan I*, et al.* Multiple layers of B cell memory with different effector functions. *Nat Immunol* **10**, 1292-1299 (2009).

25. Doria-Rose NA*, et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509**, 55-62 (2014).

26. Ellebedy AH*, et al.* Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat Immunol* **17**, 1226-1234 (2016).

27. Filias A, Theodorou GL, Mouzopoulou S, Varvarigou AA, Mantagos S, Karakantza M. Phagocytic ability of neutrophils and monocytes in neonates. *BMC pediatrics* **11**, 29 (2011).

28. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E862-870 (2015).

29. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **32**, 158-168 (2014).

30. Glanville J*, et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 20216-20221 (2009).

31. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).

32. Guinovart C*, et al.* The role of age and exposure to Plasmodium falciparum in the rate of acquisition of naturally acquired immunity: a randomized controlled trial. *PLoS One* **7**, e32362 (2012).

33. Haas A, Zimmermann K, Oxenius A. Antigen-dependent and -independent mechanisms of T and B cell hyperactivation during chronic HIV-1 infection. *Journal of virology* **85**, 12102-12113 (2011).

34. Hamid O*, et al.* Safety and tumor responses with lambrolizumab (anti-PD-

1) in melanoma. *The New England journal of medicine* **369**, 134-144 (2013).

35. Haynes BF, Kelsoe G, Harrison SC, Kepler TB. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nature biotechnology* **30**, 423-433 (2012).

36. Herati RS*, et al.* Successive annual influenza vaccination induces a recurrent oligoclonotypic memory response in circulating T follicular helper cells. *Science Immunology* **2**, (2017).

37. Horns F*, et al.* Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *eLife* **5**, (2016).

38. Hufert FT*, et al.* Germinal centre CD4+ T cells are an important site of HIV replication in vivo. *Aids* **11**, 849-857 (1997).

39. Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. *Nature* **354**, 389-392 (1991).

40. Jia Q*, et al.* Diversity index of mucosal resident T lymphocyte repertoire predicts clinical prognosis in gastric cancer. *Oncoimmunology* **4**, e1001230 (2015).

41. Jiang N*, et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine* **5**, 171ra119 (2013).

42. Jiang N, Weinstein JA, Penland L, White RA, 3rd, Fisher DS, Quake SR. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5348-5353 (2011).

43. Johnston RJ, Choi YS, Diamond JA, Yang JA, Crotty S. STAT5 is a potent negative regulator of TFH cell differentiation. *J Exp Med* **209**, 243-250 (2012).

44. Kaji T*, et al.* Distinct cellular pathways select germline-encoded and somatically mutated antibodies into immunological memory. *J Exp Med* **209**, 2079-2097 (2012).

45. Kaur K, Chowdhury S, Greenspan NS, Schreiber JR. Decreased expression of tumor necrosis factor family receptors involved in humoral

immune responses in preterm neonates. *Blood* **110**, 2948-2954 (2007).

46. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* **2**, e1501371 (2016).

47. Kim CH, Rott LS, Clark-Lewis I, Campbell DJ, Wu L, Butcher EC. Subspecialization of CXCR5+ T cells: B helper activity is focused in a germinal center-localized subset of CXCR5+ T cells. *J Exp Med* **193**, 1373-1381 (2001).

48. Kim JR, Lim HW, Kang SG, Hillsamer P, Kim CH. Human CD57+ germinal center-T cells are the major helpers for GC-B cells and induce class switch recombination. *BMC immunology* **6**, 3 (2005).

49. Kohler SL*, et al.* Germinal Center T Follicular Helper Cells Are Highly Permissive to HIV-1 and Alter Their Phenotype during Virus Replication. *J Immunol* **196**, 2711-2722 (2016).

50. Krishnamurty AT*, et al.* Somatically Hypermutated Plasmodium-Specific IgM(+) Memory B Cells Are Rapid, Plastic, Early Responders upon Malaria Rechallenge. *Immunity* **45**, 402-414 (2016).

51. Kurosaki T, Kometani K, Ise W. Memory B cells. *Nature reviews Immunology* **15**, 149-159 (2015).

52. Lefranc MP*, et al.* IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* **43**, D413-422 (2015).

53. Lin J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* **37**, 145-151 (1991).

54. Lindqvist M*, et al.* Expansion of HIV-specific T follicular helper cells in chronic HIV infection. *The Journal of clinical investigation* **122**, 3271-3280 (2012).

55. Loman NJ*, et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology* **30**, 434-439 (2012).

56. Lorenzo-Redondo R*, et al.* Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature* **530**, 51-56 (2016).

57. Machida K*, et al.* Hepatitis C virus induces a mutator phenotype: enhanced mutations of immunoglobulin and protooncogenes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4262-4267 (2004).

58. Marin AC, Gisbert JP, Chaparro M. Immunogenicity and mechanisms impairing the response to vaccines in inflammatory bowel disease. *World journal of gastroenterology* **21**, 11273-11281 (2015).

59. Matthieu P*, et al.* Follicular helper T cells serve as the major CD4 T cell compartment for HIV-1 infection, replication, and production. *J Exp Med* **210**, 143-156 (2013).

60. Michaeli M, Noga H, Tabibian-Keissar H, Barshack I, Mehr R. Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front Immunol* **3**, 386 (2012).

61. Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T cell subsets, migration patterns, and tissue residence. *Annual review of immunology* **31**, 137-161 (2013).

62. Nguyen AW*, et al.* A cocktail of humanized anti-pertussis toxin antibodies limits disease in murine and baboon models of whooping cough. *Science translational medicine* **7**, 316ra195 (2015).

63. Nussbaum C*, et al.* Neutrophil and endothelial adhesive function during human fetal ontogeny. *Journal of leukocyte biology* **93**, 175-184 (2013).

64. O'Brien PM, Tsirimonaki E, Coomber DW, Millan DW, Davis JA, Campo MS. Immunoglobulin genes expressed by B-lymphocytes infiltrating cervical carcinomas show evidence of antigen-driven selection. *Cancer immunology, immunotherapy : CII* **50**, 523-532 (2001).

65. Oestreich KJ, Mohn SE, Weinmann AS. Molecular mechanisms that control the expression and activity of Bcl-6 in TH1 cells to regulate flexibility with a TFH-like gene profile. *Nat Immunol* **13**, 405-411 (2012).

66. Pape KA, Taylor JJ, Maul RW, Gearhart PJ, Jenkins MK. Different B cell populations mediate early and late memory during an endogenous immune response. *Science* **331**, 1203-1207 (2011).

67. Parmigiani A*, et al.* Impaired antibody response to influenza vaccine in HIV-infected and uninfected aging women is associated with immune

activation and inflammation. *PLoS One* **8**, e79816 (2013).

68. Portugal S*, et al.* Malaria-associated atypical memory B cells exhibit markedly reduced B cell receptor signaling and effector function. *eLife* **4**, (2015).

69. Prabakaran P*, et al.* Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* **64**, 337-350 (2012).

70. PrabhuDas M*, et al.* Challenges in infant immunity: implications for responses to infection and vaccines. *Nature immunology* **12**, 189-194 (2011).

71. Rafael AC*, et al.* Inadequate T follicular cell help impairs B cell immunity during HIV infection. *Nat Med* **19**, 494-499 (2013).

72. Rechavi E*, et al.* Timely and spatially regulated maturation of B and T cell repertoire during human fetal development. *Science translational medicine* **7**, 276ra225 (2015).

73. Ridings J, Dinan L, Williams R, Roberton D, Zola H. Somatic mutation of immunoglobulin V(H)6 genes in human infants. *Clinical and experimental immunology* **114**, 33-39 (1998).

74. Ridings J, Nicholson IC, Goldsworthy W, Haslam R, Roberton DM, Zola H. Somatic hypermutation of immunoglobulin genes in human neonates. *Clinical and experimental immunology* **108**, 366-374 (1997).

75. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Current opinion in immunology* **25**, 646-652 (2013).

76. Schroder AE, Greiner A, Seyfert C, Berek C. Differentiation of B cells in the nonlymphoid tissue of the synovial membrane of patients with rheumatoid arthritis. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 221-225 (1996).

77. Schroeder HW, Jr., Zhang L, Philips JB, 3rd. Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood* **98**, 2745-2751 (2001).

78. Schuller SS*, et al.* Preterm neonates display altered plasmacytoid

dendritic cell function and morphology. *Journal of leukocyte biology* **93**, 781-788 (2013).

79. Shi W*, et al.* Transcriptional profiling of mouse B cell terminal differentiation defines a signature for antibody-secreting plasma cells. *Nat Immunol* **16**, 663-673 (2015).

80. Shiao YH. A new reverse transcription-polymerase chain reaction method for accurate quantification. *BMC Biotechnology* **3**, 22 (2003).

81. Shugay M*, et al.* Towards error-free profiling of immune repertoires. *Nature methods*, (2014).

82. Simon AK, Hollander GA, McMichael A. Evolution of the immune system in humans from infancy to old age. *Proceedings Biological sciences* **282**, 20143085 (2015).

83. Spitzer MH, Nolan GP. Mass Cytometry: Single Cells, Many Features. *Cell* **165**, 780-791 (2016).

84. Storm SW, Mariscalco MM, Tosi MF. Postnatal maturation of total cell content and up-regulated surface expression of Mac-1 (CD11b/CD18) in polymorphonuclear leukocytes of human infants. *Journal of leukocyte biology* **84**, 477-479 (2008).

85. Taylor JJ, Jenkins MK, Pape KA. Heterogeneity in the differentiation and function of memory B cells. *Trends in immunology* **33**, 590-597 (2012).

86. Tebas P*, et al.* Poor immunogenicity of the H1N1 2009 vaccine in well controlled HIV-infected individuals. *Aids* **24**, 2187-2192 (2010).

87. Timens W, Rozeboom T, Poppema S. Fetal and neonatal development of human spleen: an immunohistological study. *Immunology* **60**, 603-609 (1987).

88. Tipton CM*, et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nature immunology* **16**, 755-765 (2015).

89. Topalian SL, Drake CG, Pardoll DM. Immune checkpoint blockade: a common denominator approach to cancer therapy. *Cancer cell* **27**, 450-461 (2015).

90. Tran TM*, et al.* An intensive longitudinal cohort study of Malian children and adults reveals no evidence of acquired immunity to Plasmodium falciparum infection. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **57**, 40-47 (2013).

91. UNICEF. The State of the World's Children 2015: Reimagine the Future: Innovation for every child. *United Nations Children's Fund, New York, 2015*, (2015).

92. Vander Heiden JA*, et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, (2014).

93. Victora GD, Nussenzweig MC. Germinal centers. *Annual review of immunology* **30**, 429-457 (2012).

94. Vinuesa CG, Linterman MA, Yu D, MacLennan IC. Follicular Helper T Cells. *Annual review of immunology* **34**, 335-368 (2016).

95. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 13463-13468 (2013).

96. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes and immunity* **13**, 363-373 (2012).

97. Watson CT*, et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *American journal of human genetics* **92**, 530-546 (2013).

98. Weber JS*, et al.* Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial. *The Lancet Oncology* **16**, 375-384 (2015).

99. Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807-810 (2009).

100. Weinstein JS*, et al.* TFH cells progressively differentiate to regulate the

germinal center response. *Nat Immunol* **17**, 1197-1205 (2016).

101. Weisel FJ, Zuccarino-Catania GV, Chikina M, Shlomchik MJ. A Temporal Switch in the Germinal Center Determines Differential Output of Memory B and Plasma Cells. *Immunity* **44**, 116-130 (2016).

102. Weiss GE*, et al.* A positive correlation between atypical memory B cells and Plasmodium falciparum transmission intensity in cross-sectional studies in Peru and Mali. *PLoS One* **6**, e15983 (2011).

103. Weiss GE*, et al.* The Plasmodium falciparum-specific human memory B cell compartment expands gradually with repeated malaria infections. *PLoS Pathog* **6**, e1000912 (2010).

104. Weitkamp JH, Lafleur BJ, Greenberg HB, Crowe JE, Jr. Natural evolution of a human virus-specific antibody gene repertoire by somatic hypermutation requires both hotspot-directed and randomly-directed processes. *Human immunology* **66**, 666-676 (2005).

105. White MT*, et al.* A combined analysis of immunogenicity, antibody kinetics and vaccine efficacy from phase 2 trials of the RTS,S malaria vaccine. *BMC medicine* **12**, 117 (2014).

106. WHO. *World Malaria Report 2015*. Global Malaria Programme, World Health Organization (2015).

107. Wong MT*, et al.* Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell reports* **11**, 1822-1833 (2015).

108. Wong MT*, et al.* A High-Dimensional Atlas of Human T Cell Diversity Reveals Tissue-Specific Trafficking and Cytokine Signatures. *Immunity* **45**, 442-456 (2016).

109. Wu YC, Kipling D, Dunn-Walters DK. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front Immunol* **3**, 193 (2012).

110. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116**, 1070-1078 (2010).

111. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic acids research* **40**, e134 (2012).

112. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* **41**, W34-40 (2013).

113. Yu W*, et al.* Clonal Deletion Prunes but Does Not Eliminate Self-Specific alphabeta CD8(+) T Lymphocytes. *Immunity* **42**, 929-941 (2015).

114. Yu X*, et al.* Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* **455**, 532-536 (2008).

115. Zajac P, Islam S, Hochgerner H, Lonnerberg P, Linnarsson S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS One* **8**, e85270 (2013).

116. Zhang SQ*, et al.* Direct measurement of T cell receptor affinity and sequence from naive antiviral T cells. *Science translational medicine* **8**, 341ra377 (2016).

117. Zhu J*, et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 6470-6475 (2013).

118. Zhu J, Yamane H, Paul WE. Differentiation of effector CD4 T cell populations (*). *Annual review of immunology* **28**, 445-489 (2010).

119. Zinocker S*, et al.* The V gene repertoires of classical and atypical memory B cells in malaria-susceptible West African children. *IEEE Trans Vis Comput Graphics* **194**, 929-939 (2015).